

FOOD SAFETY AND DATA-DRIVEN SCIENCE:
DEVELOPMENTS USING MACHINE LEARNING AND DATABASES

A Thesis

Presented to the Faculty of the Graduate School
of Cornell University

In Partial Fulfillment of the Requirements for the Degree of
Master of Science

by

Pajau Vangay

January 2014

© 2014 Pajau Vangay

ABSTRACT

Increasing evidence suggests that persistence of *Listeria monocytogenes* in food processing plants has been the underlying cause of a number of human listeriosis outbreaks. The first part of this research study extracts criteria used by food safety experts in determining bacterial persistence in the environment, using retail delicatessen operations as a model. Using the Delphi Method, we conducted an expert elicitation with 10 food safety experts from academia, industry, and government to classify *L. monocytogenes* persistence based on environmental sampling results collected over six months for 30 retail delicatessen stores. The results were modeled using variations of random forest, support vector machine, logistic regression, and linear regression; variable importance values of random forest and support vector machine models were consolidated to rank important variables in the experts' classifications. The duration of subtype isolation ranked most important across all expert categories. Sampling site category also ranked high in importance and validation errors doubled when this covariate was removed. Support vector machine and random forest models successfully classified the data with average validation errors of 3.8% and 2.8% (n=144), respectively. Our findings indicate that (i) the frequency of isolations over time and sampling site information are critical factors for experts determining subtype persistence, (ii) food safety experts from different sectors may not use the same criteria in determining persistence, and (iii) machine learning models have potential for future use in environmental surveillance and risk management programs. Further work involving larger data sets is necessary to validate the accuracy of expert and machine classification against biological measurement of *L. monocytogenes* persistence.

To address this need for access to larger biological datasets, we developed Food Microbe Tracker, a public web-based database that allows for archiving and exchange of a variety of

molecular subtype data that can be cross-referenced with isolate source data, genetic data, and phenotypic characteristics. Data can be queried with a variety of search criteria, including DNA sequences and banding pattern data (e.g., ribotype, PFGE type). Food Microbe Tracker allows for the deposition of data on any bacterial genus and species, as well as bacteriophages and other viruses. The bacterial genera and species that currently have the most entries in this database include *Listeria monocytogenes*, *Salmonella*, *Streptococcus* spp., *Pseudomonas* spp., *Bacillus* spp., and *Paenibacillus* spp. with over 40,000 isolates present in total. The combination of pathogen and spoilage microorganism data in the database will facilitate source tracking and outbreak detection, improved discovery of emerging subtypes, and an increased understanding of transmission and ecology of these microbes. Continued addition of subtyping, genetic or phenotypic data for a variety of microbial species will broaden the database and facilitate large-scale studies on the diversity of food-associated microbes, with much potential to be extended towards the control of any organism of interest in any environment.

BIOGRAPHICAL SKETCH

Pajau Vangay was born in Longjumeau, France on November 17, 1982. She graduated from the Colorado School of Mines with a combined Bachelor and Master of Science in Mathematics and Computer Science in May of 2005. She joined Agilent Technologies in San Francisco, California, where she worked as a software engineer for five years. In 2010, Pajau returned to academia and began studying for a Master of Science degree in Food Science and Technology at Cornell University.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my family for their continued support and encouragement throughout my career. My father, Jonas Vangay, for instilling an appreciation for education from the very beginning through his own relentless pursuits in academia. My mother, Pazua Yang, for her unconditional love and support for everything I have done; and teaching me the true measurement of success. My sister, Panuly Franklin, for adding light-heartedness into my life and being a constant reminder to never take anything too seriously.

I thank Martin Wiedmann for giving me the opportunity to join the Cornell Food Safety Lab; I may have never found my niche without his support. His efforts to transition me into a life scientist always included my previous computational expertise; I will forever be indebted to him for allowing me to see my own potential. I thank him for always challenging me and keeping my best interests in mind, for my experience over the last three years has had tremendous impact on both my professional and personal philosophies moving forward.

The first study was supported by the USDA-AFRI National Integrated Food Safety Initiative Project (grant no. 2010-51110-21076). Specific appreciation is directed to the food safety experts who participated in the elicitation to make this project possible. The development of the Food Microbe Tracker database was supported by USDA Special Research Grants (2002-34459-11758; 2003-34459-12999; 2004-34459-14296), as well as funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, and the Department of Health and Human Services (Contract No. N01-A1-30054).

TABLE OF CONTENTS

BIOGRAPHICAL SKETCH	iii
ACKNOWLEDGEMENTS	iv
LIST OF FIGURES	vi
LIST OF TABLES	vii
CHAPTER 1:	
Classification of <i>Listeria monocytogenes</i> persistence in retail delicatessen environments using expert elicitation and machine learning.	1
APPENDIX	29
REFERENCES	30
CHAPTER 2:	
Food Microbe Tracker: A web-based tool for storage and comparison of food-associated microbes.	34
REFERENCES	59
SUMMARY CHAPTER	64

LIST OF FIGURES

Figure 1.1. Store 24 of the Round I questionnaire.	7
Figure 1.2. Kernel density plots of the original expert confidence values.	10
Figure 1.3. Validation errors of three models plotted with increasing months of collected sampling data.	22
Figure 2.1. A summary table generated shows an increased frequency of ribotype DUP-1044A in 1999 and 2000, representing a large human listeriosis outbreak.	43
Figure 2.2A. Table names beginning with the character A-P and their data fields.	46
Figure 2.2B. Table names beginning with the character R-Z and their data fields.	47
Figure 2.3. Automated PFGE search results search parameters.	50
Figure 2.4. Advanced search results showing differentiation among PFGE images but not ribotype images.	54

LIST OF TABLES

Table 1.1. Model comparisons using Round I results	12
Table 1.2. The set of summary statistics.	14
Table 1.3. Subtype-store instances that were difficult to classify after Round II.	15
Table 1.4. Average validation errors (%) across five runs each of three variations of the classification models using Round II results.	18
Table 1.5. Variable Importance rankings.	19
Table 1.6. Statistical results for comparison of expert group category responses.	23
Table 2.1. Data fields in Food Microbe Tracker.	38
Table 2.2. Food Microbe Tracker user levels and privileges.	40
Table 2.3. Subtype Data Summary of the Top Six Organisms in Food Microbe Tracker.	44
Table 2.4. Food Microbe Tracker query options.	49

CHAPTER 1:

Classification of *Listeria monocytogenes* persistence in retail delicatessen environments using expert elicitation and machine learning.

1.1. INTRODUCTION

Listeria monocytogenes is a Gram-positive, facultative intracellular foodborne pathogen that can cause invasive illness, listeriosis, in humans as well as animals. *L. monocytogenes* infections are estimated to be responsible for 1,600 human illnesses and 255 deaths per year in the United States⁽¹⁾. Despite having been isolated from a wide range of environments, the majority of human listeriosis cases are linked to post-processing contamination of ready-to-eat (RTE) products that can support *L. monocytogenes* growth, including RTE deli meats⁽²⁻⁴⁾. In particular, persistence of *L. monocytogenes* in processing and retail environments has been found to be a root cause of many contamination events linked to human listeriosis cases⁽⁵⁾. Multiple listeriosis outbreaks have been linked to subtypes found to be persistent in food processing environments. Specific examples involving persistent subtypes in RTE products include the 1998 outbreak linked to hot dogs, the multistate listeriosis outbreak in the 2000 attributed to turkey deli meat, and the 2008 outbreak in Canada linked to deli meats⁽⁶⁻⁹⁾. In addition to health concerns, outbreaks can result in food recalls that create significant economic burden for the food industry^(4,10). For example, the 2009 *Salmonella* outbreak in peanut butter forced the Peanut Corporation of America into bankruptcy and severely impacted other companies such as Kellogg, who reported an estimated loss of \$70 million⁽¹¹⁾. A 2010 survey conducted by the Grocery Manufacturers Association found that 77% of companies estimated their food recall costs (direct recall costs, product loss, reputation damage, etc.) to be as high as \$30 million; 23%

of companies reported even higher costs⁽¹²⁾. US food producers recalled 450,000 pounds of food due to possible contamination with *L. monocytogenes* in 2012⁽⁴⁾; in addition, the health-cost associated with listeriosis cases is estimated to be \$2.6 billion annually in the US⁽¹⁰⁾. It is clear that persistence of *L. monocytogenes* in food processing environments presents a threat to both public health as well as the economy.

Bacterial persistence in an environment is often identified by repeated isolation of a subtype over time. Previous studies regarding *L. monocytogenes* persistence used qualitative approaches in determining persistence; for example, at least three consecutive subtype isolations from the same source in a study period constituted persistence⁽¹³⁻¹⁶⁾. These qualitative definitions take on different meanings when considering the sampling plan design (e.g., sampling time period, frequency, location), subtyping method discriminatory power, overall prevalence of a subtype, likelihood of reintroduction, and other factors. Recently, Malley et al. proposed a more quantitative approach to determine *L. monocytogenes* persistence in smoked fish processing plants by comparing the distribution of the subtypes isolated during an environmental sampling study against a larger distribution of subtype isolation frequencies⁽¹⁷⁾. This study considered how common a subtype is in the environment (as represented by the larger distribution), and identified non-random isolations as persistent⁽¹⁷⁾. Similar quantitative approaches will be necessary for establishing standard, systematic methods used to identify bacterial persistence.

Expert elicitation is a process for quantifying expert opinion regarding uncertainties to address research problems in areas where traditional scientific research is infeasible or not yet available⁽¹⁸⁾. It has potential to be a reliable supplement to traditional science, and has been used successfully by U.S. government agencies in scientific areas such as environmental health, control systems, and nuclear energy⁽¹⁸⁾. The U.S. Department of Agriculture has used expert

elicitations multiple times in the past to assess risk posed to public health by various foods regulated by the Food Safety and Inspection Service⁽¹⁹⁾. In addition, expert elicitations have been used to fill data gaps in various food safety studies⁽²⁰⁻²³⁾, and a structured expert elicitation was used by Hoelzer et al. to identify cross-contamination risks and transmission pathways of *L. monocytogenes* in retail deli operations⁽²⁴⁾.

The Delphi method is one of the oldest formal expert elicitation methods; its main objective is to reach group consensus through iterative sharing of group responses⁽²⁵⁻²⁸⁾. The major steps of an iteration in the Delphi method are (i) administration of questions to participants (electronically, mail, or interview), (ii) analysis of responses and identification of responses that are in disagreement with the group consensus (as pre-defined for a given study context, generally the interquartile range), and (iii) sharing the group responses anonymously and allowing participants to revise answers⁽²³⁾. These iterations are repeated until a satisfactory level of consensus is reached as pre-defined for a given study. The advantages of the Delphi method are that it encourages knowledge sharing and convergence towards a group consensus; the disadvantages are that it may require extensive resources (time, money), and the quality of the expert opinion may degrade with too many iterations⁽¹⁸⁾. It is important to note that it is not always possible or necessary to reach consensus because disagreement among experts may actually provide significant insight.

In this study, we used the Delphi method to form food safety expert classification of *L. monocytogenes* persistence, developed quantitative models built upon these classifications, and extracted the criteria that underlie these classifications.

1.2. METHODS

1.2.1 Expert Group Selection

Fifteen food safety experts with equal representation from academia, industry, and government were conveniently identified as potential participants in this study. Details of the study objectives, logistics, and time commitment were shared with the experts before they joined the study. Of the 15 identified experts, 10 experts agreed to participate (academia, 4; industry, 3; government, 3) for the entire duration of the study. Participant identities were kept confidential throughout the study. The Cornell Institutional Review Board for Human Participants approved a formal exemption from review for this study; research activities in this study met the exemption criteria of being a confidential survey (Protocol ID#: 1210003373).

1.2.2 Questionnaire Design

The primary dataset used in the questionnaire was from an ongoing study of *Listeria monocytogenes* ecology and control in retail delicatessens^(29,30). The study sampled the environment of 30 retail stores during operation for *L. monocytogenes* and other *Listeria* species in approximately 40 environmental sample sites per store, representing food-contact, non-food-contact surfaces, and transfer points. Samples were collected once a month for six consecutive months, with positive *L. monocytogenes* isolates subtyped by pulsed-field gel electrophoresis (PFGE). Detection and subtyping of *L. monocytogenes* in this study were performed as previously described; this data set will be published separately⁽³¹⁾.

Of the original 30 stores, 12 stores were excluded from this study on the basis that during the six months, their collected samples resulted in no *L. monocytogenes* isolates, or only single isolations of distinct PFGE subtypes. To ensure a sufficiently large sample (approximated as 10 times the number of covariates)⁽³²⁾, 44 subtypes over 9 stores were generated and added to the sampling result data. These manually generated subtypes were generated either by varying the sampling sites of existing isolation patterns, or by identifying missing scenarios that were

deemed both theoretically interesting and biologically plausible, and manually generating their patterns. For example, a scenario that was not observed in our data set was the competition of multiple subtypes at a single site; this was represented as two subtypes isolated during alternating months from a single sampling location. Background noise was introduced in the generated data with the addition of 1 to 4 subtypes isolated in a single month. Novel subtypes were assigned unique PFGE types: *ApaI* types were randomly assigned values within the range in the sampling data whereas *AscI* types were randomly assigned values outside the range in the sampling data. This way, real and generated data would be indistinguishable to the experts viewing the questionnaire, but could still be filtered for analysis in our study.

The resulting dataset used for this study contained a total of 27 stores (randomly ordered for the questionnaire, with the same order distributed to all experts). Over all stores, 42 distinct PFGE subtypes created 144 unique subtype-store instances (subtype-store instance is defined as a unique instance of a given subtype in a store, e.g., CU-11-320-Store-2). Expert fatigue while filling out a lengthy questionnaire may lead to inconsistent and inaccurate responses; hence, we reduced the number of subtypes requiring expert classification. Subtypes appearing only once within a store throughout the 6-month duration were assigned the status ‘Not Persistent’ ($n = 74$) for the analysis and hence were excluded from expert classification. These excluded subtypes would be later added back into the dataset used for modeling and analysis ($n = 144$). The questionnaire developed from this process required experts to classify the 70 subtype-store instances appearing more than once as Persistent or Not Persistent (questionnaires used are available at <http://www.cals.cornell.edu/cals/foodsci/research/labs/wiedmann/links/>). A pilot study to evaluate the first questionnaire was conducted with two members of the Cornell Food Safety Lab who regularly produce and interpret PFGE data prior to distribution to the expert

panel. A full pilot study conducted with food safety experts was infeasible due to time and resource constraints; in addition, the enrollment of food safety experts into a pilot study would further limit the pool of experts available for the actual expert elicitation.

Using the Delphi Method as a guideline for administering expert elicitations, two rounds of questionnaires were developed for this study. The aim of the questionnaire design was to include as much information as possible without overwhelming the participants. The first questionnaire asked the participant to examine all 27 stores and classify subtypes that appeared more than once. A Sampling Result Table presented the 40 sampling sites (rows) over 6 months (columns) (Figure 1.1) with possible values: (a) “*L. spp*”, *Listeria* species, (b) an assigned PFGE ID such as “CU-11-320”, indicating a positive *L. monocytogenes* isolate, (c) “-”, negative results for both *L. spp* and *L. monocytogenes*, and (d) a blank, sample was not taken due to physical interference (e.g., refrigerator blocking access to a drain) or nonexistent sampling site. PFGE IDs within each store were color coded for visible contrast (i.e., sample sites that yielded a given subtype were shown in the same color), which was useful especially for stores with many subtypes. In addition to the Sampling Result Table, a Subtype Count Table was also provided to show the distribution of a given subtype among the stores (Figure 1.1). For every subtype found in a store, the number of times it was isolated within a store (“Count in Current Store”), within all 27 stores (“Total Count in All Stores”), and the number of stores it was isolated from (“Number of Stores Found In”) were provided in the Sampling Result Table. Participants were asked to respond to the questionnaire by filling in the Answer Box, which included only the subset of the subtypes isolated more than once in the store (Figure 1.1).

Store 24							
	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	
food contact	1-Pan						
	2-slicerblade	-	-	-	-	-	-
	3-slicercase						
	12-delicase	-	-	-	CU-288-464	CU-288-464	-
	13-casebymeat						
	14-delicasetray	-	-	-	-	-	-
	16-3sink_inter	-	-	-	-	-	-
	19-1sink_inter	-	-	-	-	-	-
	39-CuttingBoard	-	CU-34-413	-	-	-	Lspp
	7-rewraptable	CU-258-400	CU-258-400	-	-	CU-258-400; Lspp	CU-258-400
	40-Counter	CU-258-400; Lspp	Lspp	-	CU-258-400	Lspp	CU-258-400; Lspp
	29-coldracks	CU-258-400; Lspp	Lspp	-	CU-258-400	CU-258-400; Lspp	CU-258-400; Lspp
non food contact	35-knifejuncture	CU-258-400; Lspp	Lspp	-	Lspp	CU-258-400; Lspp	CU-258-400; Lspp
	9-dryingrack	CU-258-400	CU-258-400; Lspp	-	CU-258-400; Lspp	CU-258-400	CU-258-400; Lspp
	17-3sink_exter	Lspp	-	-	Lspp	Lspp	Lspp
	18-flrwalljunct-3b	-	-	-	-	-	-
	20-1sink_exter	-	-	-	-	Lspp	Lspp
	21-flrwalljunct-1b	-	-	-	-	-	-
	22-delidrain	-	-	-	-	-	-
	23-adjflrdm	-	-	-	-	-	-
	24-delifloor	-	-	-	-	CU-285-455	-
	26-coldfloor	-	-	-	-	-	-
	27-coldwall	-	-	-	-	-	-
	28-colddrain	-	-	-	-	-	-
transfer points	31-stdwater	-	-	-	-	-	-
	32-squeegee	-	-	-	-	-	-
	30-cartwheel	-	-	-	-	-	-
	33-hose	-	-	-	-	-	-
	34-trashcan	-	-	-	-	-	-
	36-cleaningdrain	-	-	-	-	-	-
	38-coolershelfwall	-	-	-	-	-	-
	4-slicerknob	CU-258-400; Lspp	-	-	-	CU-258-400; Lspp	-
	5-scaletop	CU-258-400; Lspp	-	-	CU-258-400	CU-258-400; Lspp	CU-258-400; Lspp
	6-scalekeys	Lspp	-	-	Lspp	Lspp	Lspp
	15-casehandle	CU-258-400; Lspp	CU-258-400; Lspp	-	CU-258-400; Lspp	CU-258-400; Lspp	CU-258-400; Lspp
	8-kniferack	-	Lspp	-	Lspp	Lspp	Lspp
	10-utensilspn	CU-258-400	CU-258-400; Lspp	-	CU-258-400; Lspp	CU-258-400; Lspp	CU-258-400; Lspp
	11-cutboard	CU-258-400; Lspp	-	-	-	-	-
	25-colddoorhand	Lspp	CU-258-400	-	-	-	-
	37-Carts	-	CU-258-400	-	-	-	-

Subtype Count Table

Subtype	Count in Current Store	Total Count in All Stores	Number of Stores Found
CU-258-400	38	44	2
CU-285-455	1	2	2
CU-288-464	2	2	1
CU-34-413	1	1	1

Answer Box

Subtype	Persistent? (Y/N)	% Confidence (1-100)
CU-258-400		
CU-288-464		

Figure 1.1. Store 24 of the Round I questionnaire. The sampling results, subtype count table, and answer box form a page in the questionnaire. Note that the Answer Box requires the expert to classify only the two subtypes isolated more than once.

Participants classified whether a subtype was persistent in a store by indicating Yes or No, accompanied by a continuous number (0-100%) representing their confidence level in their answer; to facilitate quantitative answers, experts were provided with qualitative guidance, e.g., greater than 99% was interpreted as virtually certain; greater than 90% was interpreted as extremely likely⁽³³⁾. Participants were asked to classify whether a subtype was persistent in a given store, as opposed to whether a subtype was persistent in a specific sampling site.

After we received the results from the first questionnaire, a second questionnaire was administered to form consensus on questions where experts disagreed. Similarly formatted to the first questionnaire, the second questionnaire included another column in the Answer Box that summarized how the group as a whole classified each subtype. This summary serves as the replacement for the group consensus discussion typically held between the first and second rounds in the Delphi Method. In addition, the second questionnaire was customized for each expert, prefilled with his or her previous answers from Round I for easy comparison against the group consensus. The Round II questionnaire highlighted subtypes that were in disagreement among experts in Round I, with stores reordered to match the prioritized subtypes. Although experts were required to reexamine only questions in disagreement, they had complete freedom to reevaluate and change any of their previous answers.

1.2.3 Questionnaire Preparation and Distribution

The questionnaires were prepared using Microsoft Excel 2011 (Version 14.3.2) and Adobe Acrobat Pro XI (Version 11.0.02), and were formatted as PDF Forms to ensure a consistent format of the data across operating systems. The questionnaires were distributed and collected electronically using the secure file transfer system, Cornell Dropbox (dropbox.cornell.edu) or email, according to respondent preference. Participants were given one month to complete the first questionnaire, and two weeks to complete the second questionnaire. All communications between the questionnaire administrator and the experts were electronic or by telephone; anonymity of experts was maintained among participants.

1.2.4 Data Normalization

All statistical analyses for this study were performed using R (version 2.15.2, GUI 1.53, for Mac OS X). Every classification in the questionnaires required two inputs from the experts: a

continuous confidence value (0-100) and a binary value representing Persistent (1) or Not Persistent (0). Round I responses suggested that the experts had different interpretations of the confidence value. As can be seen in Figure 1.2, Expert 7 (E7) provided confidence values between 50-100% for any subtypes classified as Persistent, and confidence values between 0-49% for any subtypes classified as Not Persistent. E7 responded with confidence values near 0% for subtypes that were sure to be Not Persistent, and near 100% for subtypes that were virtually certain to be Persistent. Experts 5, 6, and 10 (E5, E6, E10) provided confidence values between 50-100% for all classifications, regardless of whether a subtype was Persistent or Not Persistent. E5, E6, E10 confidences are synonymous with interpreting confidence as probability; confidence values near 50% are interpreted as unsure because a subtype has almost equal probability of being Persistent or Not Persistent. The remaining six experts responded with confidence values ranging between 0-100%. In order to streamline all confidence interpretations to match the majority interpretation (range 0-100%; 0% indicates low confidence and 100% indicates high confidence), qualitative follow-up questions were issued to all experts via email between Round I and Round II. The objective of these questions was to determine whether or not to rescale each expert's confidence values to 0-100%. One set of questions can be seen below:

You classified every subtype with levels of confidence greater than or equal to 50%. In order to better understand your answers, can you please clarify by replying to the following questions:

- 1.) A classification of Yes (Persistent) with 60% Confidence means:
 - a.) This subtype is Persistent, with higher than average confidence
 - b.) This subtype is Persistent, with low confidence
 - c.) Other (please explain)
- 2.) A classification of No (Not Persistent) with 50% Confidence means:
 - a.) This subtype is Not Persistent, with average confidence
 - b.) This subtype is Not Persistent, with very low confidence (equal probability of subtype being Not Persistent or Persistent)
 - c.) Other (please explain)

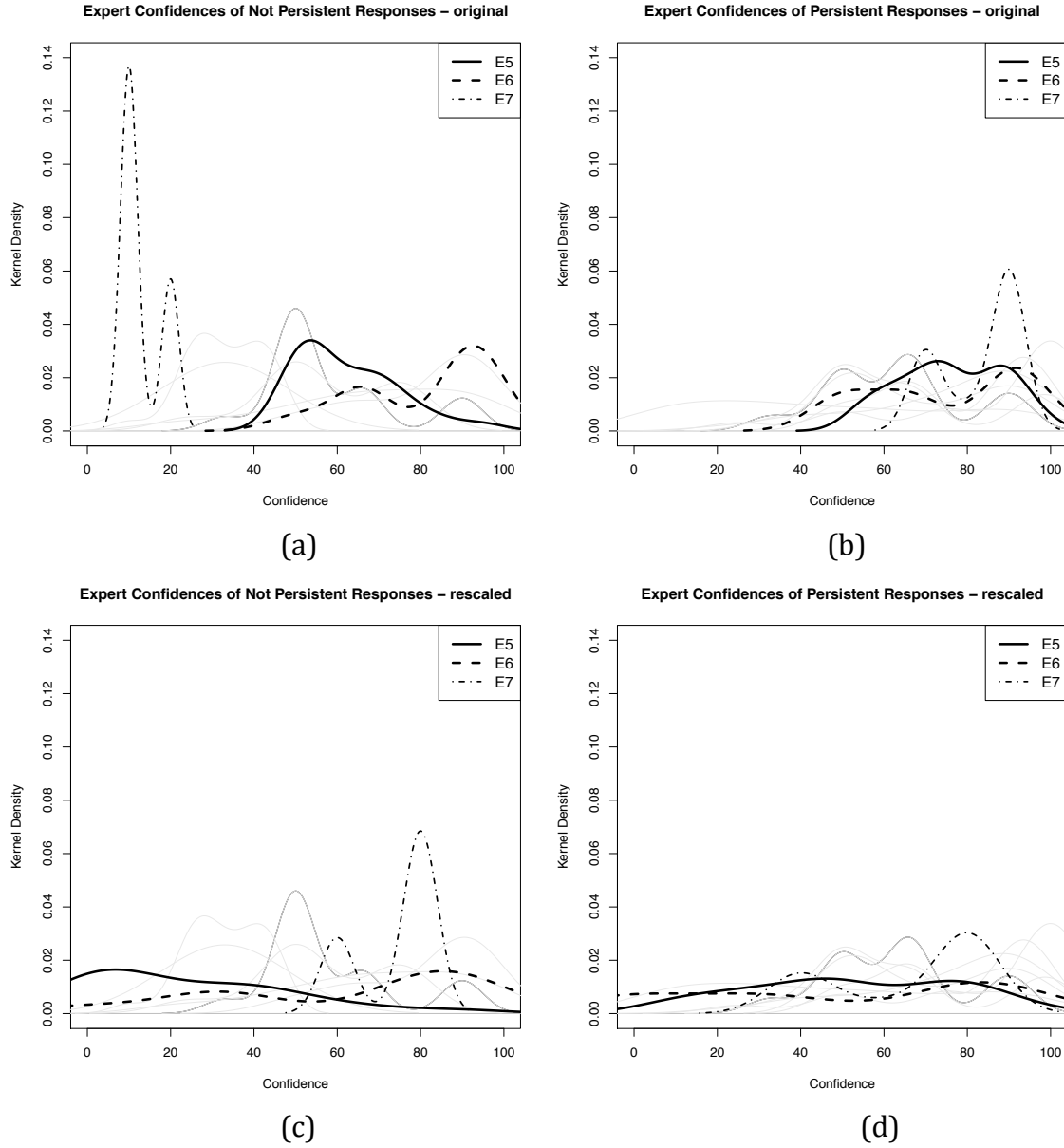


Figure 1.2. Kernel density plots of the original expert confidence values show higher confidences associated with (b) Persistent (top-right) subtypes, and lower confidences for (a) Not Persistent (top-left) subtypes. After adjusting the confidences of experts 5-7 (E5, E6, and E7) according to their qualitative feedback, confidence distributions for both (c) Not Persistent and (d) Persistent subtypes appear more evenly distributed.

Experts who did not select answer “a.)” for both questions required rescaling of their confidence values (3 experts: E5, E6, E7). Additional communication with these three experts took place over email and telephone for confirmation of the new interpretations. E5, E6, and E7’s confidence values were rescaled using the equation:

$$rescaledConfidence = 2 * | confidence - 50 |$$

For each subtype within a store, the 10 expert confidences (including the rescaled confidences) and persistence classifications were consolidated into a single value, PCRaw, in the following manner:

$$PCRaw_{subtypej} = \sum_{expert\ i=1}^{10} Confidence_{ij} * Classification_{ij} \left\{ \begin{array}{l} 1 = persistent \\ -1 = not\ persistent \end{array} \right\}$$

With 10 experts and confidences ranging from 0-100, the summary value, PCRaw, can range from -1000 to +1000. Subtypes with PCRaw values closer to -1000 are classified with high confidence as Not Persistent; PCRaw values closer to +1000 are classified with high confidence as Persistent, and PCRaw values near 0 are a) classified with low confidences (difficult to classify), or b) in non-consensus among the experts. The 74 subtypes that were presumed Not Persistent, and were excluded from classification despite being included in the elicitation, were assigned a PCRaw value of -1000.

1.2.5 Identification of Difficult to Classify Subtypes

After Round I, it was essential to identify which subtype classifications were not in consensus among the expert group for further examination in Round II. Hence, it was necessary to determine cut-offs for exactly what range of PCRaw values constituted non-consensus. Using R packages ‘kernlab’ and ‘randomForest’, Support Vector Machine (SVM) and Random Forest (RF) methods were used to fit the dataset and identify which subtypes were difficult to

classify^(34,35). The models had as input a collection of covariates (See Covariate Selection), and a binary response of Persistent or Not Persistent; subtypes with positive PCRaw values were binned as Persistent, and negative PCRaw values were Not Persistent. A leave-one-out cross-validation was employed: for each of the 144 subtypes, one subtype was left out for validation and the remaining 143 subtypes were used for training. The validation error is determined by the percentage of subtypes that were incorrectly classified over all of the 144 models. As can be seen in Table 1.1, the PCRaw values of the incorrectly classified subtypes ranged from -101 to 102 using RF (the most accurate model).

Table 1.1. Model comparisons using Round I results; minimum and maximum PCRaw values of the best performing models, SVM and RF, were used to identify subtypes that were in non-consensus and difficult to classify.

Model Comparisons	Number of Correct Classifications	Total Subtypes Classified	Validation Error	Min PCRaw Value of Misclassified Subtypes	Max PCRaw Value of Misclassified Subtypes
Three-Consecutive-Time-Periods ¹	114	144	21%	-130	916
Binomial- Counts ²	119	144	17%	-1000	-7
Binomial-Days ³	66	75	12%	-1000	131
LM ^a	130	144	9.7%	-7	916
LR ^b	132	144	8.3%	-530	804
RF ^c	141	144	2.1%	-101	102
SVM ^d	135	144	6.3%	-123	263

¹Based on the informal definition of persistence as “isolated on at least three sampling dates in a one-year period”⁽⁴⁴⁾

²Using the first binomial test as defined in Malley et al., where total subtype counts in a store were compared⁽¹⁷⁾. The significance level was chosen to minimize the validation error.

³Using the second binomial test as defined in Malley et al. ⁽¹⁵⁾. The number of days a subtype was isolated were compared, which reduced the number of subtypes to classify by removing all subtypes isolated on only one day. The significance level was chosen to minimize the validation error.

^aLinear Regression Model

^bLogistic Regression

^cSupport Vector Machine

^dRandom Forest

Using these results as guidelines, more conservative PCRaw values from -150 to 150 were chosen as the range to identify subtypes that were difficult to classify. Sixteen subtypes from

Round I of the elicitation fell within this range, and hence were highlighted for reevaluation in the Round II questionnaire.

1.2.6 Expert Category Comparisons

To assess whether there were differences in responses among the three expert categories (academia, industry, and government), two tests were performed on both Round I and Round II results. The first statistical test involved use of multiple Fisher's exact tests, comparing academia versus industry, academia versus government, and industry versus government. These three contingency tables summed the counts of Persistent and Not Persistent responses in each expert group, which were adjusted for the uneven group representation (4 experts were from academia and 3 experts each from industry and government). The second test employed logistic regression (R package 'lme4') to determine whether expert category was a significant covariate⁽³⁶⁾. This model was fitted using a logistic regression model with a fixed main effect, expert category, two random effects, subtype store ID and expert ID, with the latter nested within expert category, and an interaction between subtype store id and expert category:

$expert_response \sim (1|subtype_store_id) + expert_category + (1|expert_id/expert_category) + (1|subtype_store_id:expert_category)$. In this model, the expert response is Persistent or Not Persistent, subtype_store_id is the unique instance of a given subtype in a store (e.g., CU-11-320-Store-2), and expert category is academia, industry, or government. The expert confidence responses were used as weights in the model. The dataset used as input for the model came directly from the questionnaire (10 experts, 70 responses per expert). P-values for all tests were adjusted for multiple comparisons using the false discovery rate approach.

1.2.7 Covariate Selection

The sampling information provided in the questionnaire (Figure 1.1) was reduced to a set of 13 covariates (Table 1.2).

Table 1.2. The set of summary statistics that describe the six-month counts for subtype i found in store j , S_{ij} .

Covariate	Definition, Count of
nTotal	Total isolations of S_i over all stores
nStores	Stores S_i was isolated from
L	<i>Listeria</i> species found in store j
F	Food contact sites S_{ij} was isolated from
N	Non-food contact sites S_{ij} was isolated from
T	Transfer sites S_{ij} was isolated from
LD	Months <i>Listeria</i> species was isolated in store j
FD	Months S_{ij} was isolated from food contact sites
ND	Months S_{ij} was isolated from non-food contact sites
TD	Months S_{ij} was isolated from transfer sites
FCD	Maximum consecutive months S_{ij} was isolated from food contact sites
NCD	Maximum consecutive months S_{ij} was isolated from non-food contact sites
TCD	Maximum consecutive months S_{ij} was isolated from transfer sites

nTotal and nStores were included on the basis that the overall prevalence of a subtype could determine its persistence characteristics. The number of (non-*L. monocytogenes*) *Listeria* species isolations per store was also included; PFGE subtypes found within the same store would have the same value for *Listeria* species counts. An example of how CU-288-464 in Store 24 (Figure 1.1) is converted into numerical covariates as inputs into the models can be found in **Error!**

Reference source not found..

Table 1.3. Subtype-store instances that were difficult to classify after Round II, as defined by having PCRaw values greater than -150 and less than 150.

store	subtype	nTotal*	nStores	L	F	N	T	LD	FD	ND	TD	FCD	NCD	TCD	PCRaw
3	CU-115-583	4	1	7	0	4	0	4	0	3	0	0	1	0	124
26	CU-127-403	5	1	3	2	2	1	2	1	1	1	1	1	1	74
25	CU-308-433	2	1	5	2	0	0	1	2	0	0	2	0	0	-6
19	CU-76-461	2	1	8	0	2	0	4	0	2	0	0	2	0	-38
19	CU-146-527	2	1	8	0	2	0	4	0	2	0	0	2	0	-61
24	CU-288-464	2	1	21	2	0	0	5	2	0	0	2	0	0	-136
6	CU-182-173	2	1	0	2	0	0	0	2	0	0	1	0	0	-140

1.2.8 Statistical Models

Linear regression model (LM), logistic regression (LR), SVM, RF, and neural network were explored for fitting the dataset of this study using R packages ‘stats’, ‘kernlab’, ‘randomForest’, ‘neuralnet’, respectively^(34,35,37). All models were fit to the base equation, $response = nTotal + nStores + L + F + N + T + LD + FD + ND + TD + FCD + NCD + TCD$ (i.e., the full model) where the response is the persistence classification of a subtype consolidated from all experts. Logistic regression did not converge using all covariates and therefore was excluded from further analysis involving the full model (i.e., variable importance). For comparison among abbreviated models, logistic regression was fit to a reduced equation that excluded the number of months isolated (FD , ND , and TD) covariates. RF and SVM were constructed using the default settings; the neural network was constructed with a single hidden layer and varying number of hidden nodes. Depending on the model constraints, the response variable was either the PCRaw kept in its continuous form, or binned as a binary variable: 1 for PCRaw values between 0 and 1000, 0 for PCRaw values between -1 to -1000, inclusive (PCRaw value = 0 was arbitrarily binned as Persistent to keep the responses binary). Regardless of the response variable used, correct classifications were determined by comparing the predicted PCRaw value (binary or binned) with the binned PCRaw value of the dataset. Mean squared

errors (MSE) were calculated for models using continuous responses to provide more information on classification error without binning. With a small sample size ($n = 144$), a leave-one-out cross validation scheme was utilized for all models to ensure maximum use of the dataset during training; validation error was averaged across 5 replications for the stochastic models, SVM and RF.

1.2.9 Variable Importance Ranking

Variable importance was determined by consolidating results from random forest and support vector machine; linear regression was excluded due to its poor validation performance, logistic regression was excluded since it could not fit the full model, and neural network was not used due to its lack of a variable importance measure. RF and SVM were each trained with both binary and continuous responses, resulting in four models: 1) RF, continuous, 2) RF, binary, 3) SVM, continuous, and 4) SVM, binary. RF variable importance was determined by the percent increase in mean squared error upon random permutation of a variable and the mean increase in accuracy for continuous and binary response variables, respectively. Although variable importance is not implemented in traditional SVM algorithms, the ‘caret’ package of R provided a general computed variable importance score for SVM using the area under the receiver operating characteristic (ROC) curve, (plot of the true positive rate vs. false positive rate)⁽³⁸⁾. To consolidate all the models, the variable importance values were ranked within each model, and each covariate was averaged across all models.

1.3. RESULTS

1.3.1 Differences in Expert Elicitation Responses between Round I and Round II

In both rounds of the expert elicitation, the questionnaires contained 70 subtypes to be classified (subtypes isolated more than once out of the 144 total subtypes in the dataset); the

Round II questionnaire additionally contained the group classifications and highlighted 16 subtypes flagged for reclassification. In Round I of the expert elicitation, 43 subtypes (PCRaw mean = 546.5; PCRaw median = 643) were classified as Persistent and 27 subtypes (PCRaw mean = -269.9; PCRaw median = -205) were classified as Not Persistent. When compared with not persistent subtypes, the persistent subtypes have higher absolute mean and median PCRaw values, possibly due to the exclusion of the non-persistent subtypes appearing only once from the questionnaire. After the Round I group results were made available in the Round II questionnaire, the number of persistent subtypes decreased to 41 (PCRaw mean = 568.2; PCRaw median = 644) and not persistent subtypes increased to 29 (PCRaw mean = -315.7; PCRaw median = -256). Increases in the magnitude of the PCRaw mean of both persistent and not persistent subtypes after Round II suggest an increase in overall expert confidence. Most experts (7 out of 10) made changes only to the 16 subtypes that were highlighted for reevaluation; two experts chose not to change any of their classifications and one expert made changes to subtypes beyond the 16 highlighted. Between Round I and Round II, the PCRaw values of two subtypes changed from positive (Persistent) to negative (Not Persistent): 5 to -275 (CU-237-509, Store 13) and 74 to -6 (CU-308-433, Store 25). Note that CU-237-509 exhibits a larger change, resulting in a PCRaw value that falls outside of the difficult-to-classify range, -150 to 150.

A total of 9 subtypes exhibited large changes in PCRaw values that resulted in 7 subtypes identified as difficult to classify after Round II (Table 1.3). These 7 subtypes consist of subtypes that are isolated only once per month, over 2 or 3 months. These 7 subtypes suggest a characteristic of borderline cases that require further investigation: isolations that are too rare to confidently classify as Persistent, yet are too common to confidently classify as Not Persistent.

1.3.2 Validation Error with Machine Learning Classification

SVM and RF methods were able to classify persistence of *L. monocytogenes* subtypes in retail deli environments with accuracy 93% or better (Table 1.4).

Table 1.4. Average validation errors (%) across five runs each of three variations of the classification models using Round II results. Full Model includes all of the covariates ($response \sim nTotal + nStores + L + F + N + T + LD + FD + ND + TD + LCD + FCD + NCD + TCD$); Model without Sampling Sites excludes information regarding sampling site information ($response \sim nTotal + nStores + lspp + count + days + consDays$); Top Variables Model includes only the top three covariates as determined across the average ranking of all models ($response \sim ND + N + NCD$).

	Full Model			Model without Sampling Sites				Top Variables Model			
	LM ^a	RF ^c	SVM ^d	LM ^a	LR ^b	RF ^c	SVM ^d	LM ^a	LR ^b	RF ^c	SVM ^d
Validation Error from Binary Response Training Set	NA	2.9	6.8	NA	6.3	5.4	7.6	NA	14.5	10.4	10.4
Validation Error from Continuous Response Training Set	11.8	2.8	3.8	10.4	NA	8.5	5.1	12.5	NA	10.4	10.4

^aLinear Regression Model

^bLogistic Regression

^cSupport Vector Machine

^dRandom Forest.

When trained with continuous response variables, SVM was able to correctly classify persistence with a validation error of 3.8% and a Mean Squared Error of 2%; SVM trained with binary responses resulted in a validation error of 6.8%. RF classified with better validation error of 2.8% (and Mean Squared Error of 1.1%) when trained using continuous responses, and validation error of 2.9% when trained with binary responses. Recall that the subtypes difficult to classify had been defined as having PCRaw values between -150 and 150; hence using the binary responses, a subtype with PCRaw = -7 would be binned as Not Persistent, equating it to a subtype with PCRaw = -1000. Despite this loss of information, training with binary responses did not considerably decrease the accuracy of RF. For comparison, LM performed considerably

worse than the machine learning methods, with a validation error of 11.8%. The dataset was also fitted with a neural network model, but the validation error remained near 8% despite increasing the number of hidden nodes (in a single hidden layer) to 20. Difficulty in improving the neural network validation error with changes to other parameters (e.g., number of layers, k-fold cross-validation scheme) as well as the lack of support for a variable importance measure resulted in the exclusion of the neural network model from this study.

1.3.3 Variables of Importance Across Multiple Models

Combined variable importance values (from SVM and RF methods) ranked the most influential covariate as the number of months isolated from non-food-contact sites (ND). This covariate was consistently ranked at the top among all expert groups and over both rounds of the elicitation, with slightly larger difference in score between the first and second most important variables (Table 1.5).

Table 1.5. Variable Importance rankings averaged from support vector machine and random forest models using Round I and Round II results of the expert elicitation. Rank 1 indicates the most important variable; numerical values in parentheses represent the average ranking score across all models (identical numerical scores are interpreted as ties).

Rank	Round I				Round II			
	A ^a	I ^b	G ^c	All ^d	A ^a	I ^b	G ^c	All ^d
1	ND (12.75)	ND (13)	ND (12.75)	ND (12.75)	ND (12.75)	ND (13)	ND (12.75)	ND (13)
2	NCD (11.25)	FD (11.25)	N (11.25)	N (11)	NCD (12)	N (10.75)	N (11.25)	N (11.25)
3	N (11)	N (10.75)	NCD (10.25)	NCD (10)	N (10)	FD (10.5)	NCD (10.5)	NCD (10.25)
4	FCD (10)	NCD (10)	F (10)	FD (9.75)	FCD (10)	NCD (10.25)	F (9)	F (9.25)
5	FD (8.25)	F (9.25)	FD (10)	F (9.5)	FD (9.75)	F (9.75)	FD (8)	FD (9.25)
6	F (7.25)	nTotal (7.75)	FCD (7.75)	FCD (9.5)	F (7.25)	FCD (8)	TD (7.75)	FCD (8.75)
7	TCD (7.25)	FCD (7.75)	nTotal (7.25)	TD (7)	TD (7.25)	nTotal (7.25)	FCD (7.75)	TD (7.5)
8	TD (6.75)	TD (6.25)	TD (6)	nTotal (6)	TCD (6.25)	TD (6.25)	nTotal (7)	nTotal (5.75)
9	T (6.5)	TCD (4.75)	TCD (5)	TCD (5)	T (5.75)	T (5)	T (5.5)	TCD (5.25)
10	nTotal (4)	T (4)	T (4.75)	T (4.5)	nTotal (4)	TCD (4)	TCD (5.5)	T (4.75)
11	L (2.5)	nStores (2.25)	nStores (2)	nStores (2)	L (2.25)	nStores (2.25)	LD (2.25)	LD (2.25)
12	nStores (2)	L (2.25)	L (2)	L (2)	nStores (2)	L (2)	nStores (2)	nStores (2)
13	LD (1.5)	LD (1.75)	LD (2)	LD (2)	LD (1.75)	LD (2)	L (1.75)	L (1.75)

^aAcademia experts

^bIndustry experts

^cGovernment experts

^dAcademia, industry, and government experts

*See Table 1.2 for full descriptions of variables.

The ratio of the total number of subtypes found in non-food-contact sites, food-contact sites, and transfer points in the dataset used in this study is approximately 4:2:1, respectively; further analyses is required to determine if the high rankings of the non-food-contact site covariates are directly due to the greater number of isolates, or some underlying difference in risk information conveyed by site category positives. This uneven ratio of *L. monocytogenes* subtype distribution over site categories was left as-is in order to maintain the dataset's representation of the real world.

Multiple versions of statistical models were executed to gain insight on how the covariates affected validation error (Table 1.1). The number of consecutive months isolated and the number of isolations have been used previously to quantify persistence of *L. monocytogenes* using binomial-based statistical methods⁽¹⁷⁾. When those two covariates were used to classify this study's dataset the validation error was 12% when using the significance level that minimized the validation error (Table 1.1)⁽¹⁷⁾. When sampling site category information is removed from the model, (i.e. consolidating ND, FD, TD into days; NCD, FCD, TCD into consDays; and N, F, T into count) and trained with continuous responses, SVM and RF validation errors increased from 3.8% to 5.1% and 2.8% to 8.5%, respectively (Table 1.4). With only three consolidated covariates, both SVM and RF still performed better than the previously described binomial-based method. When only the top three most important variables, ND, N, and NCD were included as an attempt at a simplified parameterization, the validation errors increased considerably in all models (Table 1.4). The lowest validation errors found over the three variations of models in this study were obtained by training a random forest with the full set of

covariates and continuous responses (validation error of 2.8%) (Table 1.4), suggesting a complex analytical framework is required. To address any biases that the manually generated data may have introduced into the models, variable importance and validation errors were recalculated using only the original sampling results collected. Supplemental Table 1.1 shows that the variable importance rankings are similar to the models constructed with the larger dataset that included the nine manually generated stores; Supplemental Table 1.2 shows an increase in validation error, which may be attributable to the use of a smaller sample.

An attempt was made to identify the minimum number of months necessary to determine persistence by calculating model errors using data from Month 1 and 2, and iteratively including up to 6 consecutive months (Figure 1.3). As expected, all models perform better with more historical information; validation errors decreased by ~25% for the worst performing model, LM, and by ~75% for the best model, RF, when extending the data collection period from 2 months to 6 months (Figure 1.3). While the validation error appears to reach a minimum for LM, the validation errors continue to decrease for the RF and SVM, suggesting that additional sampling information may improve classification accuracy.

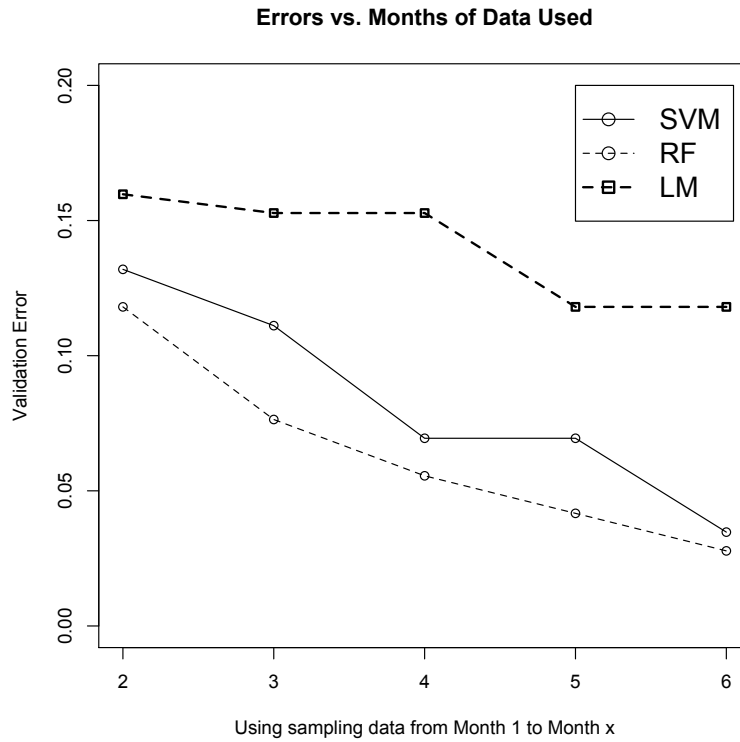


Figure 1.3. Validation errors of three models plotted with increasing months of collected sampling data. SVM = support vector machine, continuous responses; RF = random forest, continuous responses; LM = linear regression, continuous responses.

3.4 Differences in Expert Group Classification of Persistence

Using Round I results, experts in Academia classified persistent subtypes significantly different from experts in Industry (LR p-value = $8.76E-06$, Fisher's exact p-value = $1.35E-05$) and significantly different from experts in Government (LR p-value = $1.15E-03$, Fisher's Exact p-value = $1.25E-01$) (Table 1.6).

Table 1.6. Statistical results for comparison of expert group category responses using logistic regression and Fisher's exact test. P-values were adjusted for multiple comparisons using false discovery rate, and values < 0.05 are denoted in bold. A = Academia; I = Industry; G = Government.

		Round 1	Round 2
Logistic	<i>A vs. G</i>	1.15E-03	4.62E-03
Regression	<i>A vs. I</i>	8.76E-06	1.02E-05
p-values	<i>G vs. I</i>	1.75E-01	6.82E-02
Fisher's	<i>A vs. G</i>	1.25E-01	1.73E-01
exact test	<i>A vs. I</i>	1.35E-05	2.92E-04
p-values	<i>G vs. I</i>	3.50E-03	1.76E-02

In addition, based on the ratio of persistent classifications per group, Academic experts were the most conservative in assigning persistence to subtypes (44%), followed by Government (64%) then Industry experts (86%). Despite access to a summary of all responses in the second round of the elicitation, the expert group responses remained significantly different (Table 1.6). These analyses suggest there may be differences in how expert groups weigh evidence when determining persistence. This claim is limited by our expert panel size (n=10) and the fact that demographic characteristics were not collected from any of the experts; hence differences in level of education, amount of training in statistics and molecular biology, years of experience in food safety, etc. was not considered.

1.4. DISCUSSION

Our study aimed to identify whether expert opinion of *L. monocytogenes* subtype persistence in retail delicatessen environments could be learned by statistical models; if so, we were interested in determining the underlying factors that govern this classification. Our findings indicate that (i) the frequency of isolations over time in non-food-contact sites is a critical factor in subtype persistence, (ii) food safety experts from different sectors may use different criteria in determining persistence, and (iii) machine learning models have potential for future use in environmental surveillance and risk management programs.

1.4.1. Subtype persistence is dictated by the number of isolations over time and sampling site category.

The high ranking of the number of months a subtype was isolated from a non-food-contact site (ND) is consistent with the general definition of persistence as being multiple isolations over time and the fact that niche locations that harbor bacterial growth tend to be sites that are difficult to clean (usually non-food-contact sites, e.g., floors)⁽¹⁴⁾. The considerable decreases in the accuracies of the models when excluding sampling site information indicate that sampling site information is a crucial factor in determining persistence (Table 1.4).

Binomial-based statistical methods⁽¹⁷⁾ that excluded sampling site information in the number of months isolated or the number of isolations also classified this study's dataset poorly (Table 1.1). These conclusions agree with a previous review that argues environmental characteristics are critical determinants of bacterial persistence⁽¹⁴⁾, and suggest that experts already incorporate that knowledge into their decision making process.

Information on how common a subtype might be in the environment, which may provide insight on whether a subtype was being reintroduced into the store, (represented by the total number of isolations over all stores, nTotal, and the total number of stores the subtype was found in, nStores) ranked low among covariate importance. In this study, experts were advised that repeated reintroduction of a subtype from external sources was not to be considered persistence within a store. Previous research has hypothesized that common subtypes (subtypes with high nTotal and nStores values) that are widely distributed among multiple stores would give less evidence of persistence when compared with multiple isolates of very rare subtypes^(17,39). Our models gave low ranks to these subtype distribution covariates, which indicates that the experts either did not make use of (due to difficulty in interpretation, data overload, or fatigue in

participation) or gave little weight to the Subtype Count Table. Despite this information, including only the top three ranked variables in the models resulted in poor accuracy, suggesting that lower ranked variables may still be contributing to the final persistence classification.

It is important to note that these variable importance rankings may not be generalizable due to the small sample size of our dataset and the use of manually generated data. Although our use of manually generated data did not have a considerable effect on the variable importance rankings (Supplemental Table 1.1), it is important to note that there exist potential pitfalls associated with the addition of generated data. For example, the high number of *L. monocytogenes* isolates in non-food-contact sampling sites in the original dataset may be an important contributor to it also being ranked highly. If additional data was generated to balance the number of *L. monocytogenes* isolations across all sample site categories, variable importance results may be biased towards the generated data. More importantly, equal distribution of *L. monocytogenes* across all sampling site categories may not be representative of the real-world, and hence would result in models that would not be generalizable. Careful consideration should be given to the methods used in data generation and simulation.

1.4.2. Associated economic risk may be a factor in expert opinion of persistence

Experts in industry were the most willing to assign persistence to subtypes, followed by experts in government, then academia. Possible reasons for the difference in behavior may be due to associated economic risk. *L. monocytogenes* persistence in processing and retail environments represents a considerable risk for finished product contamination and associated safety, economic, and image loss (e.g., recalls, outbreaks)⁽⁴⁰⁾. For industry, there is an asymmetric cost associated with false positive and false negative results: preventative measures, such as more rigorous cleanings, are much less costly than the, sometimes irreparable,

implications of food recalls and foodborne illness outbreaks. In addition, with recent changes to food safety regulations such as the 2011 Food Safety Modernization Act (FSMA)⁽⁴¹⁾, the U.S. Food and Drug Administration (FDA) have the authority to place bracketed recalls based on findings of a very few numbers of repeated observations of identical PFGE subtypes, further justifying the defensive identification of persistence in industry. Both industry and government experts may be more sensitive to the economic or public health implications of repeated contamination, and therefore are agnostic towards considering repeated reintroduction, internal replication, or false positives as sufficient justification for classifying a subtype as not persistent. Our findings align well with a previous study that found that food safety industry and government expert opinions tend to be clustered together, and are highly influenced by their place of employment, demographic characteristics, and professional opinions⁽²⁴⁾.

There are several limitations to our study regarding differences in expert groups. The small sample size of experts within each group limits the strength of our conclusions. As previously mentioned, other factors beyond expert group such as education, training, and work experience may also provide evidence for differences among expert classifications. This point is apparent in the different interpretations and hence, the subsequent adjustments of the confidence response values among two experts in academia and one expert in government; these interpretations may be attributed to differences in statistical background. Experts may also have considerable experience in multiple sectors, which may not be well represented by their current position. Future work involving a larger expert panel and collection of demographic characteristics is necessary to evaluate differences among expert classifications of bacterial persistence.

1.4.3. Machine learning models can accurately reproduce expert opinion and have potential for integration into environmental monitoring programs and risk management decisions.

Using support vector machine and random forest models, we are able to reproduce (with 96% and 97% accuracy, respectively) expert classification of *L. monocytogenes* persistence in retail deli environments. These models can provide a systematic, reproducible method for triggering action on positive bacterial sampling results. If the food processing industry approaches to remediation are to be employed (e.g., Seek and Destroy⁽⁴²⁾), they may benefit from systematic rather than ad hoc triggers. For example, the quantitative model developed by Malley et al. was able to flag *L. monocytogenes* subtypes that were persistent, allowing for targeted interventions that successfully eradicated a subtype in at least one niche location⁽¹⁷⁾. These models also have application in the cost-benefit analyses of food safety with the potential to define the minimum thresholds necessary for determining persistence (similar to the preliminary attempt found in our study). Hence, these models may assist management decisions when evaluating the tradeoffs between increasing monitoring to obtain more accurate information and taking more immediate, less justified actions⁽⁴³⁾. It is important to note that since machine-learning models are trained to specific datasets, future work to obtain larger datasets with increased sampling frequencies and period as well as validation of persistence is necessary to strengthen the models.

1.5. CONCLUSIONS

In conclusion, we utilized expert elicitation methods to classify *L. monocytogenes* persistence based on subtyping data and implemented machine-learning models that can accurately reproduce expert opinion. Despite some differences in how experts from different sectors classify persistence, repeated isolation from non-food-contact sites over time is the main

determinant of *L. monocytogenes* persistence. These initial efforts show the value in using expert elicitation to fill data gaps, as well as the strength of machine learning methods as a quantitative decision-making tool. Future efforts involving confirmation of persistence, perhaps through a longitudinal multi-phase study, would develop a more robust training set and model that would be able to infer true bacterial persistence. Integration of quantitative approaches into bacterial monitoring systems and risk management decision-making will be essential to eradicate persistent *L. monocytogenes* subtypes, with the potential to be extended towards the control of any pathogenic organism in any environment.

APPENDIX

Supplemental Table 1.1. Variable importance ranking obtained from random forest and support vector machine models excluding manually generated data. Results are from Round II of the elicitation and include responses from all experts.

Rank	Covariate
1	ND (12.75)
2	N (11.25)
3	NCD (11)
4	FD (9)
5	nTotal (8.75)
6	F (8.75)
7	FCD (8)
8	TD (5.25)
9	nStores (5)
10	T (4.25)
11	TCD (3.5)
12	LD (2)
13	L (1.5)

Supplemental Table 1.2. Validation error obtained from fitting models to Round II results excluding manually generated data; includes all of the covariates ($response \sim nTotal + nStores + L + F + N + T + LD + FD + ND + TD + LCD + FCD + NCD + TCD$).

^aLinear Regression Model

^bSupport Vector Machine

^cRandom Forest.

	Validation Error from Binary Response Training Set	Validation Error from Continuous Response Training Set
LM ^a	NA	4.0%
RF ^b	2.0%	4.0%
SVM ^c	6.0%	6.0%

REFERENCES

1. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, Jones JL, Griffin PM. Foodborne illness acquired in the United States--major pathogens. *Emerg. Infect. Dis.* 2011;17(1): 7-15.
2. Pradhan AK, Ivanek R, Grohn YT, Bukowski R, Geornaras I, Sofos JN, Wiedmann M. Quantitative risk assessment of listeriosis-associated deaths due to *Listeria monocytogenes* contamination of deli meats originating from manufacture and retail. *J. Food Prot.* 2010;73(4): 620-30.
3. Endrikat S, Gallagher D, Pouillot R, Hicks Quesenberry H, Labarre D, Schroeder CM, Kause J. A comparative risk assessment for *Listeria monocytogenes* in prepackaged versus retail-sliced deli meat. *J. Food Prot.* 2010;73(4): 612-9.
4. USDA FSIS Recall Summaries 2012. Available at: <http://www.fsis.usda.gov/wps/portal/fsis/topics/recalls-and-public-health-alerts/recall-summaries/recall-summaries-2012>, 2013.
5. Sauders BD, Mangione K, Vincent C, Schermerhorn J, Farchione CM, Dumas NB, Bopp D, Kornstein L, Fortes ED, Windham K, Wiedmann M. Distribution of *Listeria monocytogenes* molecular subtypes among human and food isolates from New York State shows persistence of human disease--associated *Listeria monocytogenes* strains in retail environments. *J. Food Prot.* 2004;67(7): 1417-28.
6. Mead PS, Dunne EF, Graves L, Wiedmann M, Patrick M, Hunter S, Salehi E, Mostashari F, Craig A, Mshar P, Bannerman T, Sauders BD, Hayes P, Dewitt W, Sparling P, Griffin P, Morse D, Slutsker L, Swaminathan B. Nationwide outbreak of listeriosis due to contaminated meat. *Epidemiol. Infect.* 2006;134(04).
7. Olsen SJ, Patrick M, Lane K, Lee I, Wiedmann M, Huang AJ. Multistate Outbreak of *Listeria monocytogenes* Infection Linked to Delicatessen Turkey Meat. *Clin. Infect. Dis.* 2005;40(7): 962-7.
8. Nocera D, Bannerman E, Rocourt J, Jatton-Ogay K, Bille J. Characterization by DNA restriction endonuclease analysis of *Listeria monocytogenes* strains related to the Swiss epidemic of listeriosis. *J. Clin. Microbiol.* 1990;28(10): 2259-63.
9. Lessons Learned: Public Health Agency of Canada's Response to the 2008 Listeriosis Outbreak. Available at: <http://www.phac-aspc.gc.ca/fs-sa/listeria/2008-lessons-lecons-eng.php>, 2013.
10. Hoffmann S, Batz MB, Morris JG Jr. Annual cost of illness and quality-adjusted life year losses in the United States due to 14 foodborne pathogens. *J. Food Prot.* 2012;75(7): 1292-302.

11. Steenhuysen J, Dorfman B. Peanut Costs Adding Up for Food Companies. Available at: <http://www.reuters.com/article/2009/02/05/us-usa-salmonella-idUSTRE5146KL20090205>, Accessed on October 1, 2013.
12. Grocery Manufacturers Association: Capturing Recall Losses. Available at: http://www.gmaonline.org/file-manager/images/gmapublications/Capturing_Recall_Costs_GMA_Whitepaper_FINAL.pdf, Accessed on October 1, 2013.
13. Autio T, Keto-Timonen R, Lunden J, Bjorkroth J, Korkeala H. Characterisation of persistent and sporadic *Listeria monocytogenes* strains by pulsed-field gel electrophoresis (PFGE) and amplified fragment length polymorphism (AFLP). Syst. Appl. Microbiol. 2003;26(4): 539-45.
14. Carpentier B, Cerf O. Review: Persistence of *Listeria monocytogenes* in food industry equipment and premises. Int. J. Food Microbiol. 2011;145(1): 1-8.
15. Dauphin G, Ragimbeau C, Malle P. Use of PFGE typing for tracing contamination with *Listeria monocytogenes* in three cold-smoked salmon processing plants. Int. J. Food Microbiol. 2001;64(1-2): 51-61.
16. Lunden JM, Autio TJ, Sjoberg AM, Korkeala HJ. Persistent and nonpersistent *Listeria monocytogenes* contamination in meat and poultry processing plants. J. Food Prot. 2003;66(11): 2062-9.
17. Malley TJ, Stasiewicz MJ, Grohn YT, Roof S, Warchocki S, Nightingale K, Wiedmann M. Implementation of statistical tools to support identification and management of persistent *Listeria monocytogenes* contamination in smoked fish processing plants. J. Food Prot. 2013;76(5): 796-811.
18. Expert Elicitation Task Force White Paper. U.S. Environmental Protection Agency 2011.
19. Karns SA, Muth MK, Coglaiti MC. Results of an Additional Expert Elicitation on the Relative Risks of Meat and Poultry Products. 2007;53-3A94-03-12.
20. Cross P, Rigby D, Edwards-Jones G. Eliciting expert opinion on the effectiveness and practicality of interventions in the farm and rural environment to reduce human exposure to *Escherichia coli* O157. Epidemiol. Infect. 2012;140(4): 643-54.
21. Parker JS, Wilson RS, LeJeune JT, Rivers L, Doohan D. An expert guide to understanding grower decisions related to fresh fruit and vegetable contamination prevention and control. Food Control 2012;26(1): 107-16.
22. Hoffmann S, Fischbeck P, Krupnick A, McWilliams M. Using expert elicitation to link foodborne illnesses in the United States to foods. J. Food Prot. 2007;70(5): 1220-9.
23. Knol AB, Slottje P, van der Sluijs JP, Lebret E. The use of expert elicitation in environmental health impact assessment: a seven step procedure. Environ. Health 2010;9: 19,069X-9-19.

24. Hoelzer K, Oliver HF, Kohl LR, Hollingsworth J, Wells MT, Wiedmann M. Structured expert elicitation about *Listeria monocytogenes* cross-contamination in the environment of retail deli operations in the United States. *Risk Anal.* 2012;32(7): 1139-56.
25. Amara RC, Lipinski AJ. Some views on the use of expert judgment. *Technol. Forecast. Soc.* 1971;3: 279-89.
26. Rowe G, Wright G. The Delphi technique as a forecasting tool: issues and analysis. *Int. J. Forecasting.* 1999;15(4): 353.
27. Steinert M. A dissensus based online Delphi approach: An explorative research tool. *Technol. Forecast. Soc. Change* 2009;76(3): 291-300.
28. Dalkey N. An experimental study of group opinion: The Delphi method. *J. Fam. Theory Rev.* 1969;1(5): 408-26.
29. Simmons C, Wright E, Malley T, Stasiewicz M, Warchock S, Kause J, Akingbade D, Wiedmann M, Oliver HF. Ecology of *Listeria monocytogenes* in the retail deli environment. 2011.
30. Stasiewicz MJ, Wright E, Warchock S, Roof SE, Topper J, Banner N, Wiedmann M, Oliver HF. A Longitudinal Study of *L. monocytogenes* Control in Retail Delis Evaluating ATP Fluorescence as a Monitoring Tool. 2011.
31. Hoelzer K, Fortes ED, Roof SE, Wiedmann M, Sauders BD, Sanchez MD, Olsen PT, Pickett MM, Mangione KJ, Rice DH, Corby J, Stich S, Grohn YT, Oliver HF. Prevalence, distribution, and diversity of *Listeria monocytogenes* in retail environments, focusing on small establishments and establishments with a history of failed inspections. *Food Prot. Trends.* 2011;74(7): 1083-95.
32. Cohn D, Tesauro G. How Tight Are the Vapnik-Chervonenkis Bounds? *Neural Comput.* 1992;4(2): 249-69.
33. Solomon S. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge; New York: Cambridge University Press, 2007.
34. Karatzoglou A, Smola A, Hornik K. kernlab - An S4 package for kernel methods in R. 2013;0.9-18.
35. Liaw A, Wiener M. Classification and Regression by randomForest. *R News* 2012;4.6-7(3).
36. Bates D, Maechler D, Bolker B. Package 'lme4'. 2013;0.999999-2: 1-36.
37. Günther F, Fritsch S. neuralnet: Training of Neural Networks. *The R Journal* 2012;1.32(1).
38. Kuhn M. caret: Classification and Regression Training. 2013;5.17-7.

39. Ferreira V, Wiedmann M, Teixeira P, Stasiewicz MJ. *Listeria monocytogenes* persistence in food associated environments: epidemiology, strain characteristics, and implications for public health. J. Food Prot. 2013.
40. Sauders BD, Mangione K, Vincent C, Schermerhorn J, Farchione CM, Dumas NB, Bopp D, Kornstein L, Fortes ED, Windham K, Wiedmann M. Distribution of *Listeria monocytogenes* molecular subtypes among human and food isolates from New York State shows persistence of human disease-associated *Listeria monocytogenes* strains in retail environments. J. Food Prot. 2004;67(7): 1417-28.
41. Full Text of the Food Safety Modernization Act (FSMA). U. S. Food and Drug Administration 2011;111-353.
42. Malley TJ, Butts J, Wiedmann M. The Seek and Destroy Strategy: *Listeria monocytogenes* Process Controls in the Ready-to-Eat (RTE) Meat and Poultry Industry. J. Food Prot. (Submitted June 2013).
43. Ivanek R, Gröhn YT, Tauer LW, Wiedmann M. The cost and benefit of *Listeria monocytogenes* food safety measures. Crit. Rev. Food Sci. Nutr. 2004;44(7-8): 7-8.
44. Ragimbeau C, Bohin JP. Caractérisation des populations de *Listeria monocytogenes* dans la filière poisson fumé. Université des sciences et technologies de Lille. 2002.

CHAPTER 2:

Food Microbe Tracker: A web-based tool for storage and comparison of food-associated microbes.

2.1. INTRODUCTION

Large data sets on the genetic and phenotypic diversity of microorganisms causing infectious diseases can often provide new insights into infectious disease transmission, ecology, evolution, and pathogenesis. For example, Bisharat et al.⁽⁵⁾ used multilocus sequence typing (MLST) data of more than 150 *Vibrio vulnificus* isolates to characterize the origin of a new emergent human pathogenic *V. vulnificus* strain in Israel; specifically, this highly virulent new strain was shown to have evolved through hybridization of genomes from two distinct *V. vulnificus* populations. Similarly, subtype and source data for more than 2,100 *Mycobacterium tuberculosis* isolates⁽¹⁷⁾ provided evidence that different *M. tuberculosis* lineages associated with specific geographical areas are adapted to human populations originating from the respective regions. A number of web-based databases (e.g., the MLST database at www.mlst.net) allow exchange of molecular subtype data and facilitate large-scale phylogenetic studies⁽¹⁾. While large subtype databases have shown considerable promise and utility, many studies would be further facilitated by databases that integrate data for multiple subtyping methods as well as detailed geographical and source data. Unfortunately, most of the currently available subtype databases are focused on or limited to a single variety of subtype data, such as MLST or pulsed-field gel electrophoresis (PFGE) data⁽⁴²⁾.

Large subtype datasets for pathogens are also critical for surveillance studies, including outbreak detection and source tracking of foodborne diseases. For example, the PulseNet system, which was initially implemented in the US and is based on PFGE subtyping of isolates representing selected bacterial pathogens causing foodborne disease (e.g., *Listeria monocytogenes*, *Salmonella enterica*, *E. coli* O157:H7), has allowed for considerable improvement in detection of foodborne disease outbreaks and their sources⁽⁴⁴⁾. PulseNet

represents one of the few database systems that allows for deposition and sharing of banding pattern-based subtyping data and has been expanded to an international system ⁽⁴³⁾.

Unfortunately, the PulseNet database is not publicly accessible, limiting its use to public health agencies and few collaborators. A number of public databases allow for deposition of sequence based-subtype data (e.g., GenBank), but few resources for open exchange of banding pattern-based subtype data currently exist. As a result, while most scientific journals require deposition in public databases of sequence data reported in publications, no such requirements exist for the deposition of banding patterns-based data, despite the broad use of banding pattern-based subtyping methods such as PFGE, ribotyping, etc. ⁽³⁷⁾.

Although molecular subtyping methods have been traditionally applied to studies on microbial pathogen biology, these methods are also crucial for understanding the ecology, diversity, and transmission of spoilage organisms. For example, Huck et al. ⁽²²⁾ used sequence-based subtyping methods to characterize 385 *Bacillus* and *Paenibacillus* spp. isolates obtained from raw milk and pasteurized milk in two plants. Their results suggested that these spoilage organisms could be introduced in the finished product through both the raw milk as well as through post-heat treatment microbial contamination. In another study, a strong correlation between ribotypes and enzyme activities was established among the highly diverse strains of *Pseudomonas* spp. ⁽¹⁴⁾. These extracellular enzyme activities are major contributors to the spoilage and degradation of sensory qualities in milk, suggesting that specific ribotypes are more capable of causing spoilage than others ⁽¹⁴⁾. Ribotype data have also facilitated identification of specific *Pseudomonas* sources linked to spoilage problems in fluid milk and cheese in two separate plants ^(32, 38). With increased efforts in ensuring microbiological quality as a means to extend shelf-life stability, the ability to exchange, compare, and manage subtype data of spoilage organism will continue to increase in importance.

In order to further enhance the use of molecular subtyping data for both basic and applied research, we describe the development of Food Microbe Tracker, a publicly accessible molecular subtype database for food-associated microbes. This database has specifically been designed to

overcome the limitations of existing databases. For example, Food Microbe Tracker provides public access to subtype data, deposition of banding pattern-based subtype data, associations between single isolates and multiple types of subtype data (e.g. ribotype, PFGE, DNA sequence, and MLST), and integration of subtype data with detailed source data and literature citations. We anticipate that broad use of this database as well as further use of the database schema and source code for other data types (e.g., subtype data for viral and parasitic pathogen or non-food-associated microorganisms) will provide for improved subtype data exchange between different entities as well as enhanced use of subtype data generated in different laboratories for research studies and surveillance. While this database was originally developed as “PathogenTracker”, it has recently been renamed to “Food Microbe Tracker” to reflect its focus on food-associated microbes. This database is available to users worldwide for the maintenance, storage, and distribution of data on their strain collections.

2.2. METHODS AND MATERIALS

2.2.1 Design of Food Microbe Tracker

The design goals for Food Microbe Tracker were to provide: (i) robust and secure data storage, (ii) worldwide public access and collaboration, and (iii) support for multiple varieties of subtype data. In order to meet these goals, the Food Microbe Tracker software includes two parts: a relational database built using Microsoft SQL Server, and a web interface. The web interface is an ASP.NET web application and runs on Microsoft Internet Information Services; the code is written in the object-oriented language C#. The web interface requires a login to validate user data integrity and security; while it is a public database, users are required to register and will be granted the appropriate level of permission by Food Microbe Tracker database administrators. A guest login that avoids registration is available but with read-only access to limited data (e.g., sensitive information such as storage, shipping, and species specific information will be hidden). Since it is a community database, each database entry includes the following information: “user ID” and “timestamp” to identify the author and track the history of the records, “access level” to enforce which levels of users can access it, and an “active flag” to

allow for the reactivation of a record to any historical point if needed. While this database does not strictly follow ISO standards, some key concepts from appropriate ISO standards have been incorporated. For example, Microsoft's T-SQL and SQL Server are both ISO compliant and represent non-primitive data types (e.g., datetime) in a standardized manner for cross-platform integration. Common database conventions as well as our own internal conventions are enforced in order to maintain consistency, particularly among primary and foreign table keys (e.g., identical names for primary and foreign keys, appending "ID" to all primary keys, etc.).

To ensure cross-platform compatibility and multiple browser support, Food Microbe Tracker is tested and supported on the three most widely used web browsers: Microsoft Internet Explorer 9.0, Mozilla Firefox 12.0, and Google Chrome 19.0 ⁽⁴¹⁾. Food Microbe Tracker is an open-source project, hence the software and database schema are available upon request.

2.2.2 Food Microbe Tracker records

There are two major types of records in Food Microbe Tracker: sample records and isolate records. A sample record includes sample source related information (e.g., "sample obtained from", "date of sample collection", "GPS coordinates", etc.). Sample records are automatically assigned a numerical ID and may be associated with one or multiple isolate records. Isolate records in Food Microbe Tracker are assigned unique identification codes that consist of a three-letter prefix that may indicate the group or project the isolate is associated with (e.g., Food Safety Lab, FSL; Quality Milk Production Services, QMP; non-Cornell affiliated outside groups, OUT; etc.), followed by a two-character alphanumeric ID representing a user, a hyphen, and a four character sequential numeric value from 0001 – 9999 (e.g., FSL A1-2345). Once created, isolate records may be updated with a variety of source and subtyping data, including phenotypic and banding pattern-based subtype data, DNA sequence data as well as links, through GenBank accession numbers, to full genome sequences that have been deposited in GenBank (Table 2.1).

Table 2.1. Data fields in Food Microbe Tracker (in alphabetical order).

Table	Data types ^a
Additional characteristics	Contains information specific to a certain species [e.g., gene presence/absence; virulence data; etc.] or genus.
Antibiotic resistance data	Antibiotic name; Kirby Bauer (KB) diameter; sensitivity based on KB [e.g., R; resistant; I; intermediate; S; sensitive]; minimum inhibitory concentration (MIC); sensitivity based on MIC [e.g., R; I; S].
DNA sequence data	Sequence type/gene; sequence; sequence feature comments; raw data files [e.g., ABI trace files].
General isolate information	Incidence ID [i.e., a numeric code assigned to each isolate record]; lab ID [i.e., a three letter code denoting the source lab for an isolate. e.g., Food Safety Lab; FSL]; previous ID; confidential previous ID; genus; species; basis of species identification; strain; genotype; serotype; project; special strain collection; catalase activity; oxidase activity; isolate obtained from; anecdotal isolate history; representative of sample [i.e., when multiple isolates of the same species from the same sample are entered into Food Microbe Tracker; “True” is entered for the isolate that will be used for subtyping; and “False” is entered for other isolates. “True” is entered for each isolate of multiple species or if multiple isolates of the same species show different molecular subtypes.]
Genome sequence data	Bioproject accession number; web URLs; sequencing technology; assembly strategy; software strategy; approximate nucleotide coverage; sequence type; sequence.
Isolate source information	Sample ID number [i.e., numeric code automatically assigned to a new sample entered into Food Microbe Tracker]; sample ID [i.e., a code generally assigned to a sample when it was collected]; sample obtained from; description; isolated category; isolated general; isolated specific; key words; key copy; comments; confidential comments; gender; age; symptoms; fatal; month of sample collection; year of sample collection; exact date of sample collection; county; state; country; GPS coordinates.
Isolate storage information	Date frozen; aliquots; freezer; tower; box; slot; preferred media.
Isolate shipping information	Date shipped; address; institution; shipped to; shipped by.
SNP typing data	SNP Typing electropherogram images [e.g., .pdf or .fsa files]; assay.
Phage characteristics ^b	Host range; genome size; plaque characteristics; stock made date; restriction pattern image [e.g., .pdf or .tif files]; electromicroscopy image [e.g., .pdf or .tif files]; PFGE image [e.g., .txt file from Bionumerics]
PFGE data	PFGE pattern name; enzyme; pattern [e.g., text file].
Phenotypic data	Type of phenotypic test [e.g., Api 50 CHB 48 hr; api <i>Listeria</i> ; api2ONE; api20Strep; api50CHB 24h; Biolog GN]; number of tests; phenotype [i.e., data can be entered as either binary codes {010 110 type format; with 1 denoting a positive reaction and 0 denoting a negative reaction} or as octal code {the +/- data in API strips are generally expressed in an octal code - with results from 3 reactions expressed as a value between 0 and 7}].

Table 2.1. (continued)

References	PubMed ID; author; year; title; journal volume; page numbers.
Ribotype data	Ribotype pattern name; subset; enzyme; pattern [e.g., text file].

^aData storage types for all fields is text or number-based; except where noted

^bAvailable only to phage isolates as denoted by isolate designation prefix FSL-SP (*Salmonella* Phage) or FSL-LP (*Listeria* Phage).

While in most cases primary data are entered in Food Microbe Tracker, in some cases, records are linked to other data, e.g., to GenBank through accession numbers for genome sequence data.

While for some data fields, a standardized nomenclature for entries has been created (and is enforced through locked pull down menus), standardization for many other data fields still needs to be defined and enforced.

Food Microbe Tracker does not include a physical centralized repository of isolates that correspond to the data entered into the database. Requests for isolates thus need to be directed to the individuals or research groups that maintain the isolates; this information is often accessible through publications that are linked to a given isolate.

2.2.3 Submission of banding pattern-based data

Most data fields in Food Microbe Tracker require users to input data as a combination of text and/or numerical characters (Table 2.1). However, users must upload specifically formatted text files to deposit banding pattern-based data (e.g., ribotype, PFGE). Food Microbe Tracker will parse and translate the text files in order to display the PFGE or ribotype data graphically as a banding pattern. The format for files encoding ribotype patterns is defined by the Riboprinter Microbial Characterization System (DuPont Qualicon, Wilmington, DE) ⁽⁸⁾. The format for files encoding PFGE patterns is defined by the Applied Maths Bionumerics software package (Applied Maths, Saint-Matins-Latem, Belgium), which makes use of a customized conversion script developed by Applied Maths (available in the Food Microbe Tracker Online Tutorial). Food Microbe Tracker also supports the batch upload of ribotype and PFGE data as an effort to reduce the overhead of editing single isolates.

2.2.4 Data entry and accession privileges

Food Microbe Tracker users are assigned one of five different user levels that provide

different data entry, accession, and verification privileges (Table 2.2). User level 0 is reserved for database administrators and provides unrestricted database access. User level 1 refers to database management team members, who can access, enter, and verify all data (except records limited to database administrators). Only user levels 1 and 2 can verify data. User level 2 provides access to all data (except records limited to database administrators) and allows data entry. User level 3 allows data entry and access to publicly available data as well as the user's own data. Level 4 users can only access to publicly available data and may not enter data; this user level is automatically assigned to unregistered users that anonymously login to the database or to users when they create their initial password and user ID.

Table 2.2. Food Microbe Tracker user levels and privileges.

User level	Data verification level accessible by user	Default verification level for data entered by user	User can verify data
0	0 – 4	2	Y
1	1 – 4	2	Y
2	1 – 4	1	N
3	3 – 4	3	N
4	4	NA ^a	N

^aNA, not applicable.

Food Microbe Tracker is a community database and hence, a curation scheme is necessary for database managers to verify the data deposited. Verification levels are assigned to isolates and are used to identify records that have been proofread and double-checked, and to control access to specific user levels. Users with verification privileges may assign isolate records one of four verification levels. Verification level 0, 1, and 3 isolate records have not been proofread or double-checked, whereas isolate records with verification level 2 and 4 have been proofread. Only level 0 users may view verification level 0 isolate records. Isolate records with verification levels 1 and 2 are only viewable by level 0 – 2 users. Isolate records with verification level 3 are viewable by the user who created the record and by level 0 – 2 users. Isolate records with verification level 4 may be viewed by anyone who logs into the database and are thus publicly available. However, to provide data confidentiality, any specific isolate record or data

field in Food Microbe Tracker may be made inaccessible for general public users. The default verification level for new records entered by users of any given level is as follows: User levels 0 and 1, verification level 2; user level 2, verification level 1; user level 3, verification level 3.

Data verification for Food Microbe Tracker currently occurs at two levels. The database host laboratory at Cornell will qualify some users, based on their publication record and technical expertise, which allows them to enter data without routine verification by database curators; this process does not include formal assessment of technical capabilities, such as certification or proficiency testing (e.g., under formal ISO standards). For users that have not been certified (user levels 2 or 3), data that have been entered are initially assigned verification levels 1 or 3 and are reviewed by the database administrators before they are made publicly available. This review does not include formal validation of all primary data (e.g., PFGE gels, sequencing electropherograms), which is similar to other large-scale databases. Food Microbe Tracker thus should not be considered a curated database and all users thus are only granted access if they agree to a disclaimer (see <http://www.foodmicrobetracker.com/login/disclaimer.aspx?from=guestlogin>), which clarifies that the data in this database are preliminary and may contain errors.

2.2.5 Online user tutorial

We have developed an online, comprehensive tutorial with examples that instructs new users how to prepare data for Food Microbe Tracker, as well as how to use its functionalities. A link to the online tutorial can be found on the Food Microbe Tracker introduction page (<http://www.foodmicrobetracker.com/PTHelp/helpindex.htm>). Researchers who would like to contribute to this open database should register and contact Martin Wiedmann (mw16@cornell.edu) to obtain appropriate data entry privileges.

2.3. RESULTS AND DISCUSSION

Over the last decade, the private sector, academia, and government agencies have generated a significant amount of molecular subtyping data. However, resulting subtype data are often underutilized due to a lack of effective data sharing mechanisms and systems that facilitate

storage of different types of subtype data and comparison of subtype data for isolates from various sources. While some currently available molecular subtyping databases, such as PulseNet⁽⁴²⁾ and the Salm-gene system⁽¹⁵⁾, allow storage and exchange of standardized molecular subtyping data (including banding pattern-based data, such as PFGE), these databases are limited by the fact that they do not allow public data entry or access. The database at www.mlst.net⁽¹⁾ is another large subtyping database that while publicly available, is limited in scope to MLST data (i.e., DNA sequence data). Consequently, banding pattern-based data generated in a number of large studies^(7, 27, 40, 45, 47) are not readily accessible to the public. Since a comprehensive molecular subtype database that allows storage and comparison of different subtype data has not been available so far, storage and exchange of microbial subtype data has been a significant challenge to food microbiologists. Here we describe the development and initial implementation of a public web-based database (Food Microbe Tracker) that addresses this challenge. While the scope of this database is currently focused on food-associated microbes, it allows for the exchange of data for any and all microbes. In addition, this database allows for cross-referencing of disparate subtype data, including DNA sequencing data and phenotypic data. A variety of search features are available that support DNA sequence-based searches as well as banding pattern-based searches (e.g., ribotype, PFGE type). This database also allows for generation of count summary tables that can be exported for further statistical analyses (Figure 2.1). Results from these and other data analyses tools described here need to be treated with caution as results are dependent on the strain datasets available in Food Microbe Tracker; which currently does not represent comprehensive and unbiased datasets for a given population (e.g., human disease cases in the US).

	human; <i>Listeria</i> ; <i>monocytogenes</i> ; USA; NY;						
Isolates	1997	1998	1999	2000	2001	2002	2003
<i>Listeria</i> ; <i>monocytogenes</i> ; DUP-1039C	2	2	1	5	1	3	4
<i>Listeria</i> ; <i>monocytogenes</i> ; DUP-1044A	5	18	6	9	4	33	9
<i>Listeria</i> ; <i>monocytogenes</i> ; DUP-1053A	0	1	9	18	1	2	3
<i>Listeria</i> ; <i>monocytogenes</i> ; DUP-1038B	2	7	6	14	11	10	2
<i>Listeria</i> ; <i>monocytogenes</i> ; DUP-1042B	4	3	15	14	9	14	9
<i>Listeria</i> ; <i>monocytogenes</i>	24	54	75	112	55	101	58

Figure 2.1. A summary table generated using the following criteria and values for columns: isolate source general, human; genus, *Listeria*; species, *monocytogenes*; Country, USA; State, New York; year, 1997 – 2003; and the following criteria and values for rows: ribotype, DUP-1039C, DUP-1044A, DUP-1053A, DUP-1038B, DUP-1042B; genus, *Listeria*; species, *monocytogenes*. This table shows an increased frequency of ribotype DUP-1044A in 1999 and 2000, representing a large human listeriosis outbreak⁽³⁴⁾.

2.3.1 An initial implementation of a novel, public web-based database for food-associated microbe data with support for molecular subtyping data

Food Microbe Tracker is a comprehensive resource for storage and exchange of molecular subtype data for researchers worldwide, with collaborators from other universities and international institutions (e.g., Texas Tech University, Mahidol University). This database currently contains source and subtype data, including banding pattern-based and DNA sequence data, for *Listeria monocytogenes*, *Salmonella*, *Streptococcus* spp., *Pseudomonas* spp., *Bacillus* spp., and *Paenibacillus* spp., with over 40,000 isolates represented in total (Table 2.3). The Food Microbe Tracker database consists of 40 data tables (Figure 2.2A, Figure 2.2B), employing a relational database structure that allows for extensibility to include emerging subtyping methods. The database also contains data characterizing phages and allows for the addition of data on viruses.

Table 2.3. Subtype Data Summary of the Top Six Organisms in Food Microbe Tracker.

Organism	Isolates	Ribotype Images	PFGE Images	Sequences
<i>Listeria monocytogenes</i>	13,255	5,053	2,513	3,618
<i>Salmonella</i>	8,175	116	1,471	2,174
<i>Streptococcus</i> spp.	4,253	771	0	561
<i>Bacillus</i> spp.	1,742	6	0	951
<i>Pseudomonas</i> spp.	1,210	216	0	14
<i>Paenibacillus</i> spp.	1,075	8	0	950

^aData summary as of May 2012. Multiple ribotype images, PFGE images, and DNA sequences may exist for a single isolate due to the use of multiple restriction enzymes or the sequencing of different genes.

Although Food Microbe Tracker allows for subtype data comparison, it is essential for potentially ambiguous subtype data (e.g., banding pattern-based data) to be generated using standardized protocols to ensure reliable data comparison between different laboratories ⁽¹⁵⁾. Standardized PFGE protocols for subtyping of *Campylobacter jejuni*, *Salmonella*, *Shigella*, *L. monocytogenes*, and *Escherichia coli* O157:H7 ⁽¹⁰⁾, as well as *Clostridium perfringens* ⁽³³⁾ and *Vibrio cholera* ⁽¹³⁾, have been developed for the PulseNet system and are publicly available. It is suggested that PFGE data entered into Food Microbe Tracker for these seven organisms be

generated using the standard PulseNet protocols. However, since universally standard protocols for many subtyping methods and organisms do not yet exist, isolates in Food Microbe Tracker can be linked to any published subtyping method for experiment replication and data comparison. Multiple isolate records can be linked to a single literature reference and multiple literature references can be linked to a single isolate record, without requiring users to re-enter information. Food Microbe Tracker also stores results from repetitions of subtyping experiments (e.g., ribotype, PFGE, DNA sequencing) to show reproducibility when an isolate has been characterized more than once by the same or independent researchers.

AccessCounter
ID
ip
accesstime

AccessLog
ACC ID
User ID
Access Time
Action
Description

Antibiotic Resistance
Antibiotic Resistance ID
Incidence ID
Antibiotic Name
code
KB Diameter
KB Classification
MIC range
MIC
MIC Classification
Updated by
Active
Timestamp
comments

AntibioticsProfile
AbProfileID
Name
Profile
UserID
timestamp

ColumnACL
Table Name
Column Name
Display Name
Display Order
Pull Down
Data Type
Read Acc
Write Acc
Checked
Description

Genome Accession
Genome Seq ID
Bioproject
Accession
active
ID

GenomeSeqTech
Sequencing Technology
Genome Seq ID
ID

Genome Sequence
Genome Seq ID
Incidence ID
Assembly Strategy
Assembly Software
Nucleotide Coverage
Sequence Type
Sequence
Updated By
Timestamp
Active

Genome URL
Genome Seq ID
URL
ID
active

Help
helpID
topic
title
keywords
helpText

Incidence
Incidence ID
Sample ID
Representative
FSL Isolate
Designation
Previous ID
Genus
Species
basis species id
Strain
Genotype
Serotype
Isolate Obtained From
Project
NYSAGM #
Updated By
TimeStamp
Access Level
Verified By
Active
Anecdotal Isolate History
Special Strain Collection 1
Special Strain Collection 2
Project 2
Project 3
Confidential
Previous ID
Catalase Activity
Oxidase Activity

Incidence2Ref
Incidence ID
Literature ID

Incidence2Shipment
Incidence ID
SHP ID
active
timestamp
updated by

Incidence2SNP
Incidence ID
SNP ID

Literature Cited
Literature ID
Author
Year
Title
Journal
Issue
Pages
Pubmed ID
TimeStamp
Updated By
Active

personalsetting
setID
name
value
type
username
default
timestamp

PFGE Image
PFGE Image ID
Incidence ID
Enzyme Name
Pattern Name
GaussianModel
PFGeline
Bands
x
y
Image
Raw Image Path
TimeStamp
Updated By
Active

pfgeimagebk
PFGE Image ID
Incidence ID
Enzyme Name
Pattern Name
GaussianModel
PFGeline
Bands
x
y
Image
Raw Image Path
TimeStamp
Updated By
Active

Phage Characteristic
PID
Phage Incidence ID
Genome Size
Plaque Characteristic
Stock Date
Restriction Image ID
Restriction Enzyme
Electroscopy Image ID
PFGE Image ID
Updated By
Active
Timestamp

Phage Host Range
Host FSL
Phage Incidence ID
ID
Active
Timestamp
Updated By

Phage Image
ID
Image
Type
TimeStamp
Updated By
Name

Phenotype
Incidence ID
Type
Value
Updated By

Figure 2.2A. Table names beginning with the character A-P and their data fields.

RAPD Image	SampleInfo	SeqPrimers	Species2Characteristic
RAPD ID	SampleInfoID	forwardPrimer	ID
Incidence ID	SampleID	reversePrimer	Characteristic ID
Primer Name	SampleInfoTypeID	sequence ID	Genus
Primer Sequence	data	seqPrimers ID	Species
GaussianModel	updatedby	active	active
Pattern Name	Timestamp	timestamp	
Updated By	active	type	Storage
TimeStamp		Sequence	STG ID
Active	SampleInfoType	Sequence ID	Incidence ID
Bands	SampleInfoTypeID	Incidence ID	Date Frozen
x	Project	Type	Aliquots
y	Characteristic	Value	Tower Number
Raw Image Path	Datatype	Updated By	Box Number
	Access Level	TimeStamp	Slot Number
	Updated By	Active	Preferred Media
	TimeStamp	accession	Updated By
	Active		TimeStamp
Ribotype Image		Shipment	Active
Ribotype Image ID	savequery	SHP ID	Freezer
Incidence ID	savequery id	Shipping Date	
Image	userid	Shipped To	Users
Enzyme	title	Status	User ID
Ribotype Name	query	Address	User Name
Ribotype Subset	time	Institution	Password
Intensity	numberofRow	Comments	FirstName
Updated By	numberofColumn	Shipped By	LastName
TimeStamp	rowQuery		Email
Raw Image Path	titleColumn	SNPTyping	Institution
Active	titleRow	File	AccessLevel
		File Name	TimeStamp
sample	searchStatus	SNP ID	IsolatePrefix
Sample ID	searchID	TimeStamp	
Sample Identifier	searchType	Active	vocabulary
Description	textStorage	Updated By	ID
Sample Obtained From	searchtimestamp	Assay	Type
GPS Coordinates			Name
GPS Confidential	SeqFeature	SpeciesCharacteristic Value	Code
Isolate Source General	Seqfeature ID	ID	Category
Isolate Source Specific	Sequence ID	Characteristic ID	
Isolate Source Very Specific	start	Incidence ID	
Key Words	end	value	
Key Copy	featuretype	Updated by	
Comments	featurename	timestamp	
Gender		active	
Age	Seqfile		
Symptoms	Seqfile ID	Species Specific Characteristic	
Fatal	Sequence ID	Characteristic ID	
Month of Isolation	filename	Name	
Year of Isolation	rawfile	Datatype	
Exact Date of Isolation	userID	Access Level	
County where Isolated	active	Updated by	
County Confidential	timestamp	timestamp	
State where Isolated		active	
Country where Isolated			
Confidential Comments			
Timestamp			
table_timestamp			

Figure 2.2B. Table names beginning with the character R-Z and their data fields.

2.3.2 Data aggregation and database search capabilities

Since many data fields in Food Microbe Tracker are text or number-based (Table 2.1), most data fields (e.g., source category, year, state) are indexed in order to quickly search the entire database. Specifically, the quick search function will perform a global search that returns isolates with any associated data matching the given text or number. For example, users may search for all bovine-associated isolates by entering “bovine” as the search parameter in a quick search. For a more specific search, the advanced search function gives the user more control in specifying exactly which data fields to search within. For example, users may search for all isolates of a certain species, from a specific source, in a specific year (e.g., *Listeria monocytogenes* isolates from human sources in the year 2008). In addition, a batch query function is provided as a way to limit the advanced search options to only a subset of isolates, as specified by a list of isolate designations (e.g., FSL V2-001, FSL V2-002, FSL V2-003, FSL V2-004), which is entered into the batch query textbox. All search results may be viewed online through the web browser or downloaded as an Excel file. Users may also query the database for isolates linked to specific references using the PubMed ID, author, title, journal, issue, or page number information as search criteria. Finally, users may quickly access an individual isolate record by its unique identifier, the isolate designation, using the “Find Entry” function (anchored in the left margin of all Food Microbe Tracker pages).

In addition to classical text searches, Food Microbe Tracker provides a variety of search options that allow users to access and query source, phenotypic, and genetic data deposited in the database, as well as perform searches against (i) banding pattern based subtypes (e.g. ribotype or PFGE patterns), (ii) DNA sequence data for selected genes, or (iii) test-specific phenotypes (e.g., Api *Listeria*, Biolog GN, etc.) (Table 2.4). Search results are presented as a list of hypertext-linked isolates, which provide quick access to the isolate’s record of detailed characteristics, such as genotypic and phenotypic features, and source information.

Table 2.4. Food Microbe Tracker query options.

Query type	Input options
Text, number, or alphanumeric-based ^a	Quick search, single or batch query by isolate ID, advanced search using one or more text or number-based fields in the database.
Reference	Pubmed ID, author, title, journal, issue, pages.
Phenotypic data	Genus, species, phenotypic characteristic type [e.g., Api 50 CHB 48 hr, api <i>Listeria</i> , api2ONE, api20Strep, api50CHB 24h, or Biolog GN], phenotype [e.g., users may input data in either a binary or octal code format].
Automated ribotype	Desired number of matches [1 – 30], genus, species, restriction enzyme used, ribotype pattern file [e.g., text file].
PFGE	Desired number of matches [1 – 30], genus, species, restriction enzyme used, PFGE pattern file [e.g., text file].
DNA sequence	Desired number of matches, genus, species, sequence type [i.e., gene], sequence.

^aTo facilitate cross-referencing of various data types, users may select which data records will be displayed as search results when using text or number-based query functions.

Over 5,600 isolates (approximately 4,800 are *L. monocytogenes*) present in this database as of May 2012 have been characterized by automated ribotyping using the restriction enzyme *EcoRI*, and over 500 isolates (where more than 400 are *Streptococcus uberis*) have been characterized using *PvuII*. Users may perform similarity searches against the ribotype patterns in our database using their own ribotype patterns (formatted in accordance with the Qualicon Automated RiboPrinter data export format) as search criteria. This search will return up to 30 of the most similar automated ribotype patterns in an alignment.

As of May 2012, over 1,100 *L. monocytogenes* isolates present in this database have been characterized by two-enzyme (i.e., *AscI* and *ApaI*) PFGE using the PulseNet standard protocol⁽¹⁹⁾. Users may perform similarity searches against the PFGE patterns in our database using their own PFGE patterns (formatted in accordance to the Applied Maths Bionumerics software package) as search criteria (Figure 2.3). Like the ribotype search discussed above, this search will return up to 30 of the most similar PFGE patterns in an alignment.

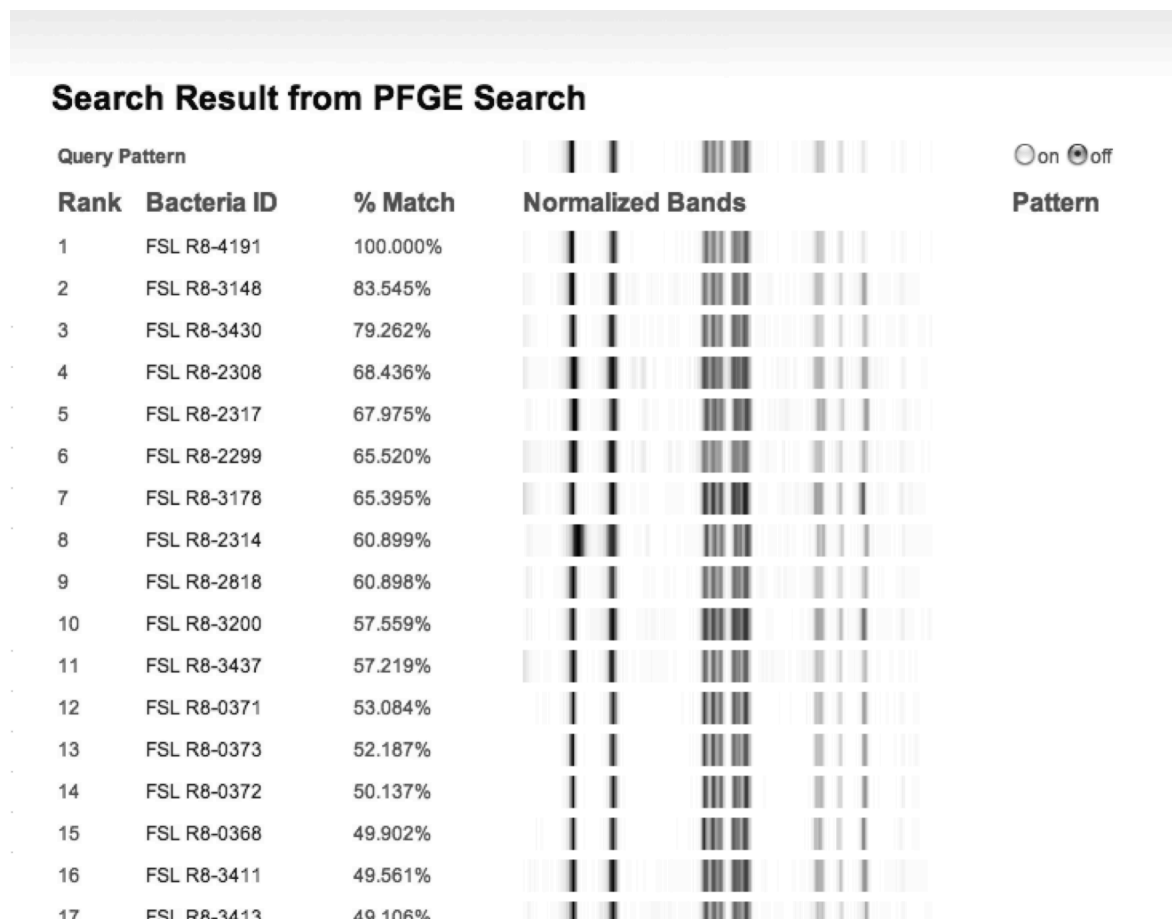


Figure 2.3. Automated PFGE search results using the following search parameters: desired number of matches: 30; PFGE pattern file: PFGE script file for isolate FSL R8-4191.

DNA sequences for the 16S rRNA gene, or select virulence or housekeeping genes for over 11,000 isolates (mainly *Salmonella* and *L. monocytogenes*) are available in our database. Food Microbe Tracker has an integrated Basic Logic Alignment Search Tool (BLAST)⁽³⁰⁾ search engine that allows users to perform similarity searches against the DNA sequences in our

database using their own DNA sequences as search criteria. Our collection also includes phenotypic array data for over 650 isolates that may be queried. For phenotypic queries, users must enter the genus and species of the organism tested, select the phenotypic characteristic type (e.g., Biolog GN, API 20 NE, API 20 Strep), and enter the phenotype (which may be in either binary or octal code).

In addition, Food Microbe Tracker allows for the creation of customized summary tables. Specifically, this function allows for generation of count summaries of isolates and their specific characteristics. For example, users may create a table showing the temporal distribution of specific human-associated *L. monocytogenes* ribotypes by selecting “*L. monocytogenes*”, “human clinical”, and specific years as parameters for columns and specific ribotypes as parameters for rows (Figure 2.1). External data utilization and statistical analyses can be achieved by exporting the summary table data to an Excel file.

The evolution of biotechnology and rapidly increasing data generation capabilities have contributed to major advancements in molecular biology, yet the rapidly changing nature of this field hinders data standardization⁽²⁹⁾. Although Food Microbe Tracker provides multiple powerful ways of retrieving data, the search results are limited by the validity and consistency of the data deposited into the database. Without existing standardization for the data fields in the database, it is not possible to programmatically validate or constrain all of the data the user may enter. For example, the nomenclature for *Salmonella* serotypes has changed immensely over time and continues to evolve, causing inconsistent naming⁽¹¹⁾. In addition, it is impractical to constrain descriptive fields such as symptoms of a patient in clinical samples, or the project where an isolate came from. With the flexibility given to users in data deposition, in order to address inconsistencies and errors, it is imperative that Food Microbe Tracker database managers serve as active data curators. Although data validation will continue to be a major effort, the growing user base of Food Microbe Tracker will facilitate future adherence to global data standards.

2.3.3 Cross-referencing of disparate subtype data, including DNA sequencing data and

phenotypic data

Unlike most currently available subtype databases, Food Microbe Tracker permits users to cross-reference disparate subtype data, including DNA sequence, phenotypic, and banding pattern-based data. Users may view multiple types of subtype data and source information for a specific set of isolates using the batch query, or for all isolates in the database using the advanced search. For either the batch query or advanced search option, users can set the criteria and display properties of each data field. For example, users may employ the advanced search to examine the diversity of PFGE types and sources associated with all *L. monocytogenes* in the database that exhibit a specific ribotype by selecting “PFGE image” and “isolated category” as data fields to be displayed, and selecting *Listeria* and *monocytogenes* as values for the genus and species criteria, respectively; and a specific ribotype pattern as a value for the ribotype name criterion. Food Microbe Tracker makes use of its ability to cross-reference disparate subtype data in order to return results that display available PFGE and source data for isolates that match the selected ribotype pattern criterion.

As sensitive and rapid subtype differentiation continues to make critical contributions to microbial food safety and food quality, subtyping methods will evolve to support various food-associated microorganisms. For example, the PulseNet PFGE protocol is the subtyping method standard for *Listeria monocytogenes* in many countries, yet ribotyping is still regularly used because of its automation and speed in attaining results^(10, 23). As a result of their differences in discriminatory power, it is crucial to be able to compare both PFGE and ribotype data of isolates⁽²³⁾. As detailed above, Food Microbe Tracker provides search capabilities that allow for side-by-side comparisons of PFGE and ribotype images, which facilitates the comparison of strains for differentiation. For example, isolates with identical ribotype patterns (e.g., ribotype DUP-1044A) may represent multiple distinct PFGE types (Figure 2.4). These types of cross-referencing results can be critical when implementing control strategies in a processing plant or when multiple subtyping methods have to be integrated for an outbreak investigation. Although the more discriminatory PFGE is the current standard subtype method for *Listeria monocytogenes*, where

historical gaps in PFGE data exist, scientists may still also value the larger repository of ribotype data in making decisions regarding persistent strains. This unique aggregation of disparate subtype data within Food Microbe Tracker is critical for leveraging the strengths of different subtyping methods, and for the transition of older to emergent subtyping methods.

Search Result

39 items found

row#	FSL	Ribotype Image	PFGE Image
1	FSL H4-004		
2	FSL F3-005		
3	FSL F3-010		
4	FSL F3-011		
5	FSL F3-015		
6	FSL F3-016		
7	FSL F3-017		
8	FSL F3-018		
9	FSL F3-019		
10	FSL F3-021		
11	FSL F3-022		
12	FSL F3-026		
13	FSL F3-027		
14	FSL F3-030		
15	FSL F3-034		
16	FSL N3-038		
17	FSL F7-068		
18	FSL F3-123		

Figure 2.4. Advanced search results using the criteria that ribotype name is DUP-1044A and PFGE Enzyme is *ApaI*, and showing ribotype Images and PFGE Images. This table shows differentiation among PFGE images but not ribotype images.

2.3.4 Food Microbe Tracker as a tool for foodborne disease outbreak investigations

Molecular subtyping has profoundly impacted public health surveillance over the last decade by allowing for improved disease cluster detection (particularly for clusters that are geographically diverse) and outbreak investigations ⁽³⁾. Molecular subtyping also facilitates phylogenetic analyses and studies on the population genetics of bacterial species ⁽³⁾. However, development of improved mechanisms and systems that facilitate storage and exchange of different subtyping and strain data for bacterial isolates from different sources may help achieve additional improvements in public health surveillance and allow broader utilization of existing subtype data in basic and applied research. While some available surveillance systems (e.g., Enter-net, Salm-gene, PulseNet), have been instrumental for the identification and control of a number of outbreaks of human disease ^(15, 18, 21, 43), there is still a need to develop improved, more comprehensive tools and systems for public health surveillance ^(24, 26, 44). Food Microbe Tracker provides a unique resource that can potentially facilitate more rapid human (and animal) outbreak detection, particularly when more users routinely deposit subtype data for isolates from clinical cases; development of automated scripts to import data from other sources (e.g., PulseNet, potentially after a certain holding phase) would facilitate this and reduce workload that would be associated with entering data into multiple databases. Users may query the database for matches to specific subtypes representing human or animal cases to determine if a specific subtype has recently been associated with human or animal disease, and request summary counts on these subtypes to determine if the incidence of disease caused by a specific subtype has abnormally increased. The use of Food Microbe Tracker to monitor the incidence of disease caused by specific subtypes may also allow for improved detection of emerging pathogen subtypes, including multi-drug resistant (MDR) clones. For example, users may compare subtype data for strains that have increased in prevalence with other strain characteristics to determine if a subtype has previously exhibited resistance to an antibiotic. Since Food Microbe Tracker

currently contains subtype and antibiotic resistance data for over 900 *Salmonella* isolates, it may be particularly useful for the identification and tracking of new MDR *Salmonella* clones.

2.3.5 Food Microbe Tracker facilitates improved control strategies of food-associated microbes

In addition to its role in improved outbreak detection, molecular subtyping methods are important for understanding the diversity of food-associated microbes in order to develop effective control strategies. For example, Manfreda et al.⁽³¹⁾ compared ribotype data for Italian Gorgonzola cheese-associated *L. monocytogenes* isolates with ribotype data in Food Microbe Tracker. These researchers discovered that a number of *L. monocytogenes* ribotypes isolated in Italy were indistinguishable from *L. monocytogenes* ribotypes for isolates from different seafood and dairy products⁽³¹⁾ and some human sporadic cases, indicating that some *L. monocytogenes* ribotypes capable of causing sporadic human disease may be common to certain types of foods. Similarly, Lyautey et al.⁽²⁸⁾ used Food Microbe Tracker to show that isolates from different types of fecal samples collected in Canada, matched ribotypes of isolates from human sporadic and epidemic listeriosis cases. Neves et al.⁽³⁵⁾ used Food Microbe Tracker to show that PFGE types for *L. monocytogenes* isolates obtained in Portugal matched PFGE types from isolates obtained from other countries allowing for identification of common and widely distributed PFGE types. Although traditionally applied to microbial pathogen biology, molecular subtyping methods have also proved successful in their applications to spoilage organisms^(14, 22). For example, Huck et al.⁽²²⁾ used an *rpoB* subtyping method to characterize *Bacillus* and *Paenibacillus* spp. isolates from raw and pasteurized milk samples from two dairy plants. They isolated unique *rpoB* allelic types in pasteurized milk that were not isolated in raw milk, indicating there were potentially persistent subtypes within the processing plants that were contributing to post-pasteurization contamination⁽²²⁾. Although the study did not examine persistence of these subtypes, Food Microbe Tracker can subsequently be used to determine whether these subtypes have been isolated from these (or any other) processing plants previously. In both of these examples, simple subtype data comparison through Food Microbe

Tracker can provide preliminary information about the distribution of subtypes, which may consequently facilitate the development of improved control strategies^(22, 31). *Bacillus*, *Pseudomonas*, and *Paenibacillus*, three important genera of spoilage organisms, are already well represented in Food Microbe Tracker (Table 2.3) and provide a starting point for the use of this database to support efforts in the control of spoilage organisms.

In addition to subtype diversity and distribution, hypothesis generation about sources and reservoirs for specific subtypes is just as important for the development of effective and specific control strategies^(3, 46). For spoilage organisms, users can compare source information of subtypes to develop hypotheses regarding their transmission into specific foods. Similarly for pathogenic organisms, users may compare subtype data for outbreak related subtypes to subtype data for isolates from a variety of sources to develop clues about possible outbreak sources. While identification of indistinguishable subtypes in isolates from human or animal clinical cases and specific sources (e.g., foods, processing plants, animals) does not necessarily prove an epidemiologic link, identification of isolates with subtypes matching outbreak strains may help in source tracking⁽⁴⁴⁾.

2.3.6 Identification of new emerging strains and strains with unique characteristics

While many of the Food Microbe Tracker's potential applications discussed relate to the analysis of pathogen and spoilage organism subtype data, Food Microbe Tracker's storage capabilities and functionalities are applicable to all bacteria and can even include phages and viruses. Specifically, it is anticipated that this database will become an increasingly valuable tool for large-scale comparison and analysis of molecular biology data and for studies on the diversity of food-associated microbes as existing subtyping, genetic or phenotypic data for a variety of microbial species are deposited. For example, Food Microbe Tracker's ability to integrate different types of subtype data for given isolates will allow users to identify links between specific molecular subtypes and meaningful biological traits, similar to the Entrez database retrieval system, which allows discovery of sequence function by facilitating comparison between genetic or amino acid sequences of unknown function with annotated data from the

major DNA (e.g., GenBank) and protein sequence databases ⁽⁴⁾. Specifically, Food Microbe Tracker can provide for the definition of subtypes where phenotypic and genetic data provide explanations as to causes of source associations, virulence differences, or unique transmission characteristics observed for these specific subtypes. For example, Nightingale et al. ⁽³⁶⁾ determined that *L. monocytogenes* isolates with ribotype DUP-1062A represent a clonal group characterized by attenuated invasiveness for human intestinal epithelial cells due to a premature stop codon in *inlA*, a *L. monocytogenes* gene critical for invasion of human intestinal epithelial cells. Searches within the Food Microbe Tracker database allowed these researchers to determine that the ribotype was more commonly associated with foods than human clinical cases, indicating that mutations in *inlA* may not only be responsible for an attenuated invasion phenotype of *L. monocytogenes* with ribotype DUP-1062A ⁽³⁶⁾, but also appear to reduce the ability of isolates with this ribotype to cause human disease.

2.4. CONCLUSIONS

In conclusion, we have developed Food Microbe Tracker, a web-based database that allows storage and exchange of different varieties of molecular subtype data (including banding pattern-based and DNA sequence data) for microbial isolates from any source. Food Microbe Tracker fills a critical gap in networks for the exchange of molecular biological data by providing a publicly available, comprehensive database of microbial source and subtype information with integrated data query and aggregation. This database represents a platform that can facilitate source tracking and increased understanding of the ecology and transmission of food-associated microbes, as well as improved disease surveillance for pathogenic organisms. In addition, this database provides a unique resource for basic and applied studies on the ecology, population genetics, and diversity of food-associated microorganisms. While we have outlined a number of potential applications of this database, Food Microbe Tracker's strength in the analysis of microbial subtype data is limited by the amount and speed of user data deposition. Members of the worldwide research community are thus encouraged to contribute their existing data to this database to allow open data exchange and facilitate large-scale analyses and studies on microbial

biodiversity. Banding pattern-based molecular subtype data as well as DNA sequence data produced by our group ^(2, 16, 20, 25, 39) are already freely available through Food Microbe Tracker for internet-based data mining. In addition, as with other non-curated molecular biology databases ^(6, 9) despite the clear value of these databases, data quality issues are a concern and innovative procedures for data curation and quality checks will need to be implemented and developed for Food Microbe Tracker (see the CODA 2000 ⁽¹²⁾ for a discussion of these issues).

REFERENCES

1. Aanensen, D. M., and B. G. Spratt. 2005. The multilocus sequence typing network: Mlst.net. *Nucleic Acids Res.* 33:W728-33.
2. Alcaine, S. D., Y. Soyer, L. D. Warnick, W. L. Su, S. Sukhnanand, J. Richards, E. D. Fortes, P. McDonough, T. P. Root, N. B. Dumas, Y. Grohn, and M. Wiedmann. 2006. Multilocus sequence typing supports the hypothesis that cow- and human-associated *Salmonella* isolates represent distinct and overlapping populations. *Appl. Environ. Microbiol.* 72:7575-7585.
3. Barrett, T. J., P. Gerner-Smidt, and B. Swaminathan. 2006. Interpretation of pulsed-field gel electrophoresis patterns in foodborne disease investigations and surveillance. *Foodborne Pathog. Dis.* 3:20-31.
4. Benson, D. A., I. Karsch-Mizrachi, K. Clark, D. J. Lipman, J. Ostell, and E. W. Sayers. 2012. GenBank. *Nucleic Acids Res.* 40:D48-53.
5. Bisharat, N., D. I. Cohen, M. C. Maiden, D. W. Crook, T. Peto, and R. M. Harding. 2007. The evolution of genetic structure in the marine pathogen, *Vibrio vulnificus*. *Infect. Genet. Evol.* 7:685-693.
6. Bork, P., and A. Bairoch. 1996. Go hunting in sequence databases but watch out for the traps. *Trends Genet.* 12:425-427.
7. Borucki, M. K., J. Reynolds, C. C. Gay, K. L. McElwain, S. H. Kim, D. P. Knowles, and J. Hu. 2004. Dairy farm reservoir of *Listeria monocytogenes* sporadic and epidemic strains. *J. Food Prot.* 67:2496-2499.
8. Bruce, J. L. 1996. Automated system rapidly identifies and characterizes microorganisms in food. *Food Technol.* 50:77-81.
9. Brunak, S., J. Engelbrecht, and S. Knudsen. 1990. Cleaning up gene databases. *Nature* 343:123.
10. Centers for Disease Control and Prevention. 2011. PulseNet protocols. Available at: www.cdc.gov/pulsenet/protocols.htm. Accessed 6 June 2006.
11. Centers for Disease Control and Prevention. 2011. National salmonella surveillance data. Available at: www.cdc.gov/ncidod/dbmd/phlisdata/salmonella.htm. Accessed 14 May 2012.
12. CODATA Task Group on biological macromolecules and colleagues. Committee on Data for Science and Technology of the International Council of Scientific Unions. 2000. Quality control in databanks for molecular biology. *Bioessays* 22:1024-1034.
13. Cooper, K. L., C. K. Luey, M. Bird, J. Terajima, G. B. Nair, K. M. Kam, E. Arakawa, A. Safa, D. T. Cheung, C. P. Law, H. Watanabe, K. Kubota, B. Swaminathan, and E. M. Ribot. 2006. Development and validation of a PulseNet standardized pulsed-field gel electrophoresis protocol for subtyping of *Vibrio cholerae*. *Foodborne Pathog. Dis.* 3:51-58.

14. Dogan, B., and K. J. Boor. 2003. Genetic diversity and spoilage potentials among *Pseudomonas* spp. isolated from fluid milk products and dairy processing plants. *Appl. Environ. Microbiol.* 69:130-138.
15. Fisher, I. S., E. J. Threlfall, Enter-net, and Salm-gene. 2005. The Enter-net and Salm-gene databases of foodborne bacterial pathogens that cause human infections in Europe and beyond: An international collaboration in surveillance and the development of intervention strategies. *Epidemiol. Infect.* 133:1-7.
16. Fugett, E. B., D. Schoonmaker-Bopp, N. B. Dumas, J. Corby, and M. Wiedmann. 2007. Pulsed-field gel electrophoresis (PFGE) analysis of temporally matched *Listeria monocytogenes* isolates from human clinical cases, foods, ruminant farms, and urban and natural environments reveals source-associated as well as widely distributed PFGE types. *J. Clin. Microbiol.* 45:865-873.
17. Gagneux, S., K. DeRiemer, T. Van, M. Kato-Maeda, B. C. de Jong, S. Narayanan, M. Nicol, S. Niemann, K. Kremer, M. C. Gutierrez, M. Hilty, P. C. Hopewell, and P. M. Small. 2006. Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U. S. A.* 103:2869-2873.
18. Gerner-Smidt, P., K. Hise, J. Kincaid, S. Hunter, S. Rolando, E. Hyytia-Trees, E. M. Ribot, B. Swaminathan, and PulseNet Taskforce. 2006. PulseNet USA: A five-year update. *Foodborne Pathog. Dis.* 3:9-19.
19. Graves, L. M., and B. Swaminathan. 2001. PulseNet standardized protocol for subtyping *Listeria monocytogenes* by macrorestriction and pulsed-field gel electrophoresis. *Int. J. Food Microbiol.* 65:55-62.
20. Gray, M. J., R. N. Zadoks, E. D. Fortes, B. Dogan, S. Cai, Y. Chen, V. N. Scott, D. E. Gombas, K. J. Boor, and M. Wiedmann. 2004. *Listeria monocytogenes* isolates from foods and humans form distinct but overlapping populations. *Appl. Environ. Microbiol.* 70:5833-5841.
21. Hedberg, C. W., and J. M. Besser. 2006. Commentary: Cluster evaluation, PulseNet, and public health practice. *Foodborne Pathog. Dis.* 3:32-35.
22. Huck, J. R., N. H. Woodcock, R. D. Ralyea, and K. J. Boor. 2007. Molecular subtyping and characterization of psychrotolerant endospore-forming bacteria in two New York state fluid milk processing systems. *J. Food Prot.* 70:2354-2364.
23. Inglis, T. J., L. O'Reilly, N. Foster, A. Clair, and J. Sampson. 2002. Comparison of rapid, automated ribotyping and DNA macrorestriction analysis of *Burkholderia pseudomallei*. *J. Clin. Microbiol.* 40:3198-3203.
24. Isaacson, R. E., M. Torrence, and M. R. Buckley. 2005. Preharvest food safety and security. Available at: <http://academy.asm.org/index.php/colloquium-program/-food-microbiology/196-preharvest-food-safety-and-security>. Accessed 18 July 2006.

25. Jeffers, G. T., J. L. Bruce, P. L. McDonough, J. Scarlett, K. J. Boor, and M. Wiedmann. 2001. Comparative genetic characterization of *Listeria monocytogenes* isolates from human and animal listeriosis cases. *Microbiology* 147:1095-1104.
26. Kathariou, S. 2002. *Listeria monocytogenes* virulence and pathogenicity, a food safety perspective. *J. Food Prot.* 64:1811-1829.
27. Lukinmaa, S., K. Aarnisalo, M. L. Suihko, and A. Siitonen. 2004. Diversity of *Listeria monocytogenes* isolates of human and food origin studied by serotyping, automated ribotyping and pulsed-field gel electrophoresis. *Clin. Microbiol. Infect.* 10:562-568.
28. Lyautey, E., A. Hartmann, F. Pagotto, K. Tyler, D. R. Lapen, G. Wilkes, P. Piveteau, A. Rieu, W. J. Robertson, D. T. Medeiros, T. A. Edge, V. Gannon, and E. Topp. 2007. Characteristics and frequency of detection of fecal *Listeria monocytogenes* shed by livestock, wildlife, and humans. *Can. J. Microbiol.* 53:1158-1167.
29. MacMullen, W. J., and S. O. Denn. 2005. Information problems in molecular biology and bioinformatics. *J. Am. Soc. Inf. Sci. Technol.* 56:447-456.
30. Madden, T. 2002. The NCBI handbook: The BLAST sequence analysis tool. Available at: <http://www.ncbi.nlm.nih.gov/books/NBK21097/>. Accessed 6 June 2006.
31. Manfreda, G., A. De Cesare, S. Stella, M. Cozzi, and C. Cantoni. 2005. Occurrence and ribotypes of *Listeria monocytogenes* in Gorgonzola cheeses. *Int. J. Food Microbiol.* 102:287-293.
32. Martin, N. H., S. C. Murphy, R. D. Ralyea, M. Wiedmann, and K. J. Boor. 2011. When cheese gets the blues: *Pseudomonas fluorescens* as the causative agent of cheese spoilage. *J. Dairy Sci.* 94:3176-3183.
33. Maslanka, S. E., J. G. Kerr, G. Williams, J. M. Barbaree, L. A. Carson, J. M. Miller, and B. Swaminathan. 1999. Molecular subtyping of *Clostridium perfringens* by pulsed-field gel electrophoresis to facilitate food-borne-disease outbreak investigations. *J. Clin. Microbiol.* 37:2209-2214.
34. Mead, P. S., E. F. Dunne, L. Graves, M. Wiedmann, M. Patrick, S. Hunter, E. Salehi, F. Mostashari, A. Craig, P. Mshar, T. Bannerman, B. D. Sauders, P. Hayes, W. Dewitt, P. Sparling, P. Griffin, D. Morse, L. Slutsker, B. Swaminathan, and *Listeria* Outbreak Working Group. 2006. Nationwide outbreak of listeriosis due to contaminated meat. *Epidemiol. Infect.* 134:744-751.
35. Neves, E., A. Lourenco, A. C. Silva, R. Coutinho, and L. Brito. 2008. Pulsed-field gel electrophoresis (PFGE) analysis of *Listeria monocytogenes* isolates from different sources and geographical origins and representative of the twelve serovars. *Syst. Appl. Microbiol.* 31:387-392.
36. Nightingale, K. K., K. Windham, K. E. Martin, M. Yeung, and M. Wiedmann. 2005. Select *Listeria monocytogenes* subtypes commonly found in foods carry distinct nonsense mutations in *inlA*, leading to expression of truncated and secreted internalin A, and are associated with

- a reduced invasion phenotype for human intestinal epithelial cells. *Appl. Environ. Microbiol.* 71:8764-8772.
37. Olive, D. M., and P. Bean. 1999. Principles and applications of methods for DNA-based typing of microbial organisms. *J. Clin. Microbiol.* 37:1661-1669.
 38. Ralyea, R. D., M. Wiedmann, and K. J. Boor. 1998. Bacterial tracking in a dairy production system using phenotypic and ribotyping methods. *J. Food Prot.* 61:1336-1340.
 39. Sauders, B. D., K. Mangione, C. Vincent, J. Schermerhorn, C. M. Farchione, N. B. Dumas, D. Bopp, L. Kornstein, E. D. Fortes, K. Windham, and M. Wiedmann. 2004. Distribution of *Listeria monocytogenes* molecular subtypes among human and food isolates from New York state shows persistence of human disease-associated *Listeria monocytogenes* strains in retail environments. *J. Food Prot.* 67:1417-1428.
 40. Sheng, H., M. A. Davis, H. J. Knecht, D. D. Hancock, J. Van Donkersgoed, and C. J. Hovde. 2005. Characterization of a shiga toxin-, intimin-, and enterotoxin hemolysin-producing *Escherichia coli* ONT:H25 strain commonly isolated from healthy cattle. *J. Clin. Microbiol.* 43:3213-3220.
 41. StatCounter. 2012. StatCounter global statistics. Available at: <http://gs.statcounter.com>. Accessed 5 June 2012.
 42. Swaminathan, B., T. J. Barrett, S. B. Hunter, R. V. Tauxe, and CDC PulseNet Task Force. 2001. PulseNet: The molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerg. Infect. Dis.* 7:382-389.
 43. Swaminathan, B., P. Gerner-Smidt, L. K. Ng, S. Lukinmaa, K. M. Kam, S. Rolando, E. P. Gutierrez, and N. Binsztein. 2006. Building PulseNet international: An interconnected system of laboratory networks to facilitate timely public health recognition and response to foodborne disease outbreaks and emerging foodborne diseases. *Foodborne Pathog. Dis.* 3:36-50.
 44. Tauxe, R. V. 2006. Molecular subtyping and the transformation of public health. *Foodborne Pathog. Dis.* 3:4-8.
 45. Vela, A. I., J. F. Fernandez-Garayzabal, J. A. Vazquez, M. V. Latre, M. M. Blanco, M. A. Moreno, L. de La Fuente, J. Marco, C. Franco, A. Cepeda, A. A. Rodriguez Moure, G. Suarez, and L. Dominguez. 2001. Molecular typing by pulsed-field gel electrophoresis of Spanish animal and human *Listeria monocytogenes* isolates. *Appl. Environ. Microbiol.* 67:5840-5843.
 46. White, D. G., P. Fedorka-Cray, and T. C. Miller. 2006. The national antimicrobial resistance monitoring systems (NARMS). Available at: <http://www.nmconline.org/articles/NARMS.pdf>. Accessed 6 June 2006.
 47. Zhao, S., P. F. McDermott, S. Friedman, J. Abbott, S. Ayers, A. Glenn, E. Hall-Robinson, S. K. Hubert, H. Harbottle, R. D. Walker, T. M. Chiller, and D. G. White. 2006. Antimicrobial

resistance and genetic relatedness among *Salmonella* from retail foods of animal origin: NARMS retail meat surveillance. *Foodborne Pathog. Dis.* 3:106-117.

SUMMARY CHAPTER

Listeria monocytogenes is the foodborne pathogen responsible for the invasive illness, listeriosis, which can further cause abortion, septicemia, and meningitis. This organism is ubiquitous in the environment, but only a few strains are capable of causing disease. Research efforts to understand the ecology and transmission pathways of *L. monocytogenes* into the food system, as well as to differentiate and characterize *L. monocytogenes* strains, are necessary for public health. The majority of human listeriosis cases have been linked to post-processing contamination of ready-to-eat deli meats and increasing evidence also suggests that persistence of *L. monocytogenes* in food processing environments is the underlying cause of many listeriosis outbreaks. To further the development of quantitative, systematic methods for identifying bacterial persistence, our efforts aimed to extract the criteria used by food safety experts in identifying persistence of *L. monocytogenes* based on environment sampling data. Future efforts to validate the accuracy of these methods will require larger datasets of bacterial sampling data obtained with highly discriminatory subtyping methods. To support this, we also sought to encourage data sharing of bacterial subtyping information on a global scale. We developed a public database to manage microbial phenotypic and genotypic characteristics, isolate source information, and most importantly, subtyping information obtained with various methods.

Our initial efforts used expert elicitation to classify *L. monocytogenes* persistence based on environmental sampling results. The classified dataset was used to construct various statistical and machine learning models as a means to extract the underlying criteria that determines bacterial persistence. As a result, our findings indicated that the frequency of isolations over time and sampling site information are critical factors in subtype persistence, and food safety experts from different sectors do not use the same criteria in determining persistence. Based on the

accuracy of the models in being able to reproduce expert opinion, there is potential for future use in environmental surveillance and risk management programs. It also may be useful for cost-benefit analyses, specifically when evaluating the tradeoffs between taking a precautionary approach to collect more data, or a more immediate action. Future work with larger datasets is necessary to validate the accuracy and scope of these models. It would also be advantageous to use large datasets of environmental sampling data that have been validated against biological measurements of *L. monocytogenes* persistence as input into these models, eliminating any noise introduced with expert opinion.

To address this need, we developed Food Microbe Tracker, a public web-based database that allows archiving and exchange of a variety of molecular subtype data that can be cross-referenced with isolate source data, genetic data, and phenotypic characteristics. This database provides the infrastructure necessary to encourage data sharing across all sectors, and has the unique capability to cross-reference subtyping data obtained with different methods. This is especially useful when comparing bacterial strains subtyped with modern methods against those subtyped with older, obsolete methods; allowing access to a potentially large set of historical data that may be critical in determining bacterial persistence and source tracking. As with any database, continued addition of subtyping, genetic and phenotypic data is necessary to facilitate data-mining efforts.

Although initially developed for food-associated microbes, Food Microbe Tracker has the capability to manage data for any bacterial genus or species, bacteriophages, and other viruses. In addition, our efforts to model *L. monocytogenes* persistence can be applied to other organisms and environments, such as the persistence of *Clostridium difficile* in hospital environments, or the persistent asymptomatic infection of *Mycobacterium tuberculosis* in the lungs. Both of these

studies represent strategies necessary to manage and make sense of the growing explosion of “big data” in scientific research.